Submitted to *Operations Research*
manuscript OPRE-2017-08-413.R1

The discrete moment problem with nonconvex shape constraints

Xi Chen
Stern School of Business, New York University, xchen3@stern.nyu.edu

Simai He, Bo Jiang
School of Information Management and Engineering, Shanghai University of Finance and Economics, simaihe@mail.shufe.edu.cn, jiangbo@mail.shufe.edu.cn

Christopher Thomas Ryan
UBC Sauder School of Business, University of British Columbia, chris.ryan@sauder.ubc.ca

Teng Zhang
Management Science and Engineering, Stanford University, tengz@stanford.edu

The discrete moment problem is a foundational problem in distribution-free robust optimization, where the goal is to find a worst-case distribution that satisfies a given set of moments. This paper studies the discrete moment problems with additional "shape constraints" that guarantee the worst-case distribution is either log-concave (LC), has an increasing failure rate (IFR), or increasing generalized failure rate (IGFR). These classes of shape constraints have not previously been studied in the literature, in part due to their inherent nonconvexities. Nonetheless, these classes are useful in practice, with applications in revenue management, reliability, and inventory control. We characterize the structure of optimal extreme point distributions under these constraints. We show, for example, that an optimal extreme point solution to a moment problem with m moments and LC shape constraints is piecewise geometric with at most m pieces. This optimality structure allows us to design an exact algorithm for computing optimal solutions in a low-dimensional space of parameters. We also leverage this structure to study a robust newsvendor problem with shape constraints and compute optimal solutions.

Key words: Robust optimization, moment problem, shape constraints, reverse convex programming
History: Original submission 1 August 2017, returned 28 November 2018 for a first revision and September 2019 for a second revision, accepted on 5 December 2019

1. Introduction

The moment problem is a classical problem in analysis and optimization, with roots dating back to the middle of the nineteenth century. At that time, the goal was to bound tail probabilities and expectations with given distributional moment information. This initial goal remains an active one to the present day. For example, Bertsimas and Popescu (2005) provide tight closed-form bounds of $\Pr(X \geq (1+\delta)\mathbb{E}X)$ given the first three moments of a random variable X. He et al. (2010) extend the problem for random variables given first-, second- and fourth-order moments.

Beyond these foundational questions, the moment problem serves as an important building block in a variety of applications in the stochastic and robust optimization literatures (Prékopa 2013, Popescu 2005, Saghafian and Tomlin 2016, Rujeerapaiboon et al. 2018, Tian et al. 2017). In particular, moment problems are foundational to distribution-free robust optimization, where insight into the structure of optimal distributions is used to devise algorithms and describe properties of optimal decisions. A classic example of this approach is due to Scarf et al. (1958), who leverage the fact that an optimal solution to the moment problem given the first two moments is a sum of two Dirac measures. This insight provides an analytical formula for the optimal inventory decision in a robust version of the newsvendor problem. A vast literature on robust optimization builds on these initial insights in a variety of facets (see, e.g., Goh and Sim (2010), Delage and Ye (2010), Jiang et al. (2012), Bandi and Bertsimas (2012), Bertsimas et al. (2018), Long and Qi (2014), Gao and Kleywegt (2016), Gao et al. (2017), Natarajan et al. (2017), Chen et al. (2018) among many others).

The focus of this paper is the discrete moment problem, an important special case of the general moment problem that is less well-studied in the literature. In the discrete moment problem, the underlying sample space is a discrete set. The focus on discrete distributions is well-justified. In operations-management applications, unknown demand distributions are often "bulky", representing, for instance, bulk orders in a supply chain (Swaminathan and Shanthikumar 1999). Banciu and Mirchandani (2013) motivate the importance of modeling demand as a discrete distribution. They cite a number of papers that model demand as continuous distributions that need to be adjusted when the realism of discrete distributions is incorporated. For example, Bagnoli et al. (1989) show how standard results in durable-goods monopolies (including the Coase conjecture) fail when demand is discretely distributed. Similar observations hold in the supply-chain literature (see, in particular, Swaminathan and Shanthikumar (1999)).

Coming back to discrete moment problems, the work of Prékopa (see, e.g., Prékopa (1990)) makes a fundamental contribution by devising efficient linear programming methods to study discrete moment problems. Prékopa and co-authors use these insights to study a variety of applications, including project management (Prékopa et al. 2004), network reliability (Prékopa and Boros 1991), and finance applications (Prékopa et al. 2016). Project management has also been studied in the robust optimization literature (see, e.g., Bertsimas et al. (2006)).

In classical versions of the moment problem (including the previously mentioned works of Prékopa and his co-authors), constraints arise from specifying a finite number of moments. One criticism of this approach is that it can result in bounds that are too weak to be meaningful, or in the case of robust optimization with only moment constraints, result in decisions that are too conservative. For instance, Scarf's solution for the newsvendor problem can suggest ordering *no* inventory even when the profit margin is relatively high (Perakis and Roels 2008). This conservatism has driven researchers to introduce additional constraints, including those on the *shape* of the distribution. For example, Perakis and Roels (2008) study the newsvendor problem by leveraging non-moment information, including symmetry and unimodality. Han et al. (2014) study the newsvendor problem by relaxing the usual assumption of risk neutrality. Saghafian and Tomlin (2016) analyze the problem with a bound on tail probabilities, and Natarajan et al. (2017) recently developed closed-form solutions under asymmetric demand information. In all cases, more intuitive and less conservative inventory decisions result. Other robust optimization papers that consider shape constraints include Li et al. (2017), who study the chance constraints and conditional value-at-risk constraints when the distributional information consists of the first two moments and unimodality, Lam and Mottet (2017), who study tail distributions with convex-shape constraints, and Hanasusanto et al. (2015), who study the multi-item newsvendor problems with multimodal demand distributions.

Introducing shape constraints brings new theoretical challenges. A seminal paper by Popescu (2005) provides a general framework for studying continuous moment problems under shape constraints. Her framework captures symmetry and unimodality constraints, among others. These moment problems are formulated as semi-definite programs (SDPs) that are polynomial-time solvable. For the discrete moment problem, we are aware of only one paper (Subasi et al. 2009) that considers shape constraints. Subasi et al. (2009) adapt Prékopa's linear programming (LP) methodology to include unimodality, which is modeled by an additional set of linear constraints.

Both Popescu (2005) and Subasi et al. (2009) illustrate how a class of constraints can be adapted into existing computational frameworks, SDP-based in the case of Popescu (2005) and LP-based in the case of Subasi et al. (2009). However, important shape constraints of practical significance remain that do not naturally fit into these settings. In this paper, we focus on three shape constraints: log-concavity (LC), increasing failure rate (IFR), and increasing generalized failure rate (IGFR) for discrete distributions (defined in Section 2 below). Here, we briefly highlight the importance of each class of distributions.

(i) LC distributions arise naturally in many applications. For example, Subasi et al. (2009) illustrate how the length of a critical path in a PERT model, where beta distributions describe individual task times, has an LC distribution but its other properties (other than moments inferred by the beta distributions) are unknown. Log-concavity has a wide range of applications to statistical modeling and estimation (Walther 2000). For instance, Duembgen et al. (2011) show how log-concavity allows the estimation of a distribution based on arbitrarily censored data (which is a common form of data for demand observations). Log-concavity also plays a critical role in economics (Bagnoli and Bergstrom 2005). For example, in contract theory, one commonly assumes an agent's type is an LC random variable (Laffont and Tirole 1988). The log-concavity of a distribution function has also been widely used in the theory of regulation (Baron and Myerson 1982, Lewis and Sappington 1988), and in characterizing efficient auctions (Myerson and Satterthwaite 1983, Matthews 1987).

(ii) IFR distributions play an important role in reliability theory (Barlow and Proschan 1996), inventory management (Gavirneni et al. 1999), and contract theory (Dai and

Jerath 2016). The IFR property is often useful for simplifying optimality conditions to facilitate the derivation of properties of optimal decisions that yield managerial insights. One reason for the prevalence of IFR distributions in applications is that they are closed under sums of random variables (and the associated convolutions of distribution functions).

(iii) IGFR distributions are prevalent in the theory of revenue management and supply chain management (see Lariviere (2006), Ziya et al. (2004), Banciu and Mirchandani (2013) and the references therein), where it is typically assumed to ensure expected revenue is unimodular in the choice of price or quantity. Banciu and Mirchandani (2013) pointed out that the definition of IGFR in the case of discrete distributions needs to be adapted from the continuous case to ensure this unimodularity. We use their definition in this paper (see Definition 3). Banciu and Mirchandani (2013) also show many of the most common discrete distributions used in applications (see Table 2 in that paper) satisfy the IGFR property.

The above references motivate LC, IFR, and IGFR distributions generally, either their continuous or discrete versions, although continuous distributions appear more frequently in the literature. However, solving a moment problem with continuous shape constraints is technically challenging and beyond the scope of this paper. We hope the techniques developed in this paper (especially a connection to reverse convex optimization) shed light on the continuous version in future work. Second, discrete distributions find important applications. Examples include discrete IFR in reliability systems (Chakraborty 2015) and discrete IGFR in modeling demand (Banciu and Mirchandani 2013). Moreover, given data points sampled from a discrete distribution, statistical tests have been well developed for testing whether the underlying discrete distribution is an IFR distribution (Sudheesh et al. 2015).

In Section 2, we show the standard characterizations of discrete LC, IFR, and IGFR distributions, when added to the moment problem, make the resulting problem nonconvex and thus not amenable to either an SDP or LP formulation. Indeed, when Subasi et al. (2009) derive LC distributions in their applications, they *relax* the LC property to unimodality, a shape constraint that *can* be approached by LP techniques.

Given this nonconvex problem, one could turn to approximation methods, including conic-optimization techniques. The copositive cone and its dual are well known to be powerful tools to convert nonconvex problems equivalently into convex ones (see, e.g., Natarajan et al. (2011), Peña et al. (2015), Xu and Burer (2018), Burer (2009), Hanasusanto and Kuhn (2018)). Natarajan et al. (2011) consider a moment problem without shape constraints, using copositive cones in representing an underlying nonconvexity due to integrality. Tools needed to directly connect copositive cones to modeling nonconvexities due to shape constraints can be found in Peña et al. (2015), although the authors do not explicitly note this connection. We show below that the LC discrete moment problem we consider here can be cast as a polynomial optimization problem (see formulation (2) in Section 2.1) that can be further written as a completely positive conic problem. Despite the convexity of completely positive conic problems, they typically remain computationally intractable. Further relaxation is often required to obtain an approximate solution.

We do not follow an approximation approach. The nonconvexities that arise in our problems are of a certain type that can be leveraged to provide an exact global optimization algorithm and analytical results on the structure of optimal solutions. Indeed, the feasible regions have *reverse convex* properties (as introduced in Meyer (1970) and later developed in Hillestad and Jacobsen (1980) among others). A set is reverse convex if its complement is convex. Reverse-convex programming is a little-studied field that has largely found application in the global optimization literature (see, for e.g., Horst and Thoai (1999)). To our knowledge, this theory has not been leveraged in robust optimization.

In Section 3, we extend standard results in the reverse convex programming literature (in particular, those of Hillestad and Jacobsen (1980)) so that they apply to our setting, by introducing the notion of reverse convexity *relative* to another set. The main benefit is that we can show reverse-convex programs of this type have the following appealing structure: there exist optimal extreme point solutions with a basic feasible structure analogous to basic feasible solutions in linear programming. The basic feasible structure reveals (in Section 4) that optimal extreme point distributions in the LC, IFR, and IGFR settings have a piecewise geometric structure. Our analytical characterization allows us to solve these discrete moment problems as low-dimensional systems of polynomial equations. We propose a specialized computation scheme for working with such systems, which allows us to provide numerical bounds on probabilities that are tighter than those in the existing literature, including those bounds that leverage unimodal shape constraints (see Section 5). In Section 5.3, we further leverage this structure to numerically study a robust newsvendor problem, assuming demand satisfies an IGFR distribution.

Due to space constraints, technical proofs are found in the e-companion when not provided in the main text.

Summary of contributions

The main focus of the paper is on the theoretical properties of LC-, IFR-, and IGFR-constrained moment problems, where we provide structural results on optimal solutions. For the LC case, we show optimal solutions exist that are piecewise geometric, and for the IFR and IGFR cases, we show the tail probabilities of optimal distributions are piecewise geometric.

Our structural results and computational approaches suggest a wide range of applications due to the prevalence of these classes of shape-constraints in real applications, as discussed above. For instance, our results can provide new bounds on tail inequalities (i.e., $\Pr(X \geq a)$) for a random variable X under moment and shape constraints. We provide a numerical framework for computing these bounds. Of particular interest is showing the applicability of these methods in robust optimization. We take the robust newsvendor problem as an illustrative example. We show the optimal structure of solutions to discrete moment problems proves useful in providing sufficient structure for the inner minimization problems of max-min robust optimization formulations, greatly enabling the solution of the outer maximization problem. This result builds on the recent paper Ninh et al. (2019) that looks at the newsvendor problem under discrete demand distributions by including shape constraints.

Finally, we prove a new result on a generalized form of reverse convex optimization (Theorem 2) that may be of independent interest, with potential applications to other nonconvex optimization problems. We believe the connection between reverse convexity (in particular, the extreme point characterization of such problems) and the structure of discrete distributions is also of interest for those who study optimization under uncertainty. At a high level, our method shows that when a distribution is defined by local conditions (roughly a single property at each outcome of the distribution) then many of these constraints will be tight at an extremal distribution. Setting these tight constraints to equalities then lends useful insights about the structure of optimal extremal solutions. In what follows, we give several instances that make this high-level reasoning precise.

Notation

We use the following notation throughout the paper. Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}^n$ the vector space of n-dimensional real vectors. Moreover, let $\mathbb{R}^n_+$ denote the set of n-dimensional vectors with all nonnegative components and let $\mathbb{R}^n_{++}$ denote the set of n-dimensional vectors with all positive components. The closure of the set S in $\mathbb{R}^n$ (in the usual topology) is denoted $\mathrm{cl}(S)$ and its boundary by $\mathrm{bd}(S)$. Let $\mathbb{E}[\cdot]$ denote the expectation operator and $\mathbb{1}_A$ the indicator function of set A; that is, $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise.

Let $[k, \ell] = \{k, k+1, \ldots, \ell-1, \ell\}$ denote the set of consecutive integers, starting with integer k and ending with integer ℓ. Similarly, let $(k, \ell) = \{k+1, k+2, \ldots, \ell-2, \ell-1\}$. To avoid confusion, we do not use $[\cdot, \cdot]$ and $(\cdot, \cdot)$ in their usual sense as intervals in $\mathbb{R}$. For k, j positive integers, $\binom{k}{j}$ denotes the binomial coefficient of k choose j; that is, it counts the number of ways to choose subsets of size j from k objects.

2. The discrete moment problem with nonconvex shape constraints

We study the discrete version of the classical problem of moments with m moments (cf. Popescu (2005)):

$$\max_{\mu \in \mathcal{P}} \int_{\Omega} f(w) d\mu$$

where $\mathcal{P}$ is a subset of measures μ on a discrete measurable space Ω (with elements denoted by w), $\mathcal{B}$ is a σ-algebra on Ω, f is a measurable function on Ω, and $q_i \in \mathbb{R}$ for $i \in [0, m]$. We take $q_0 = 1$ to ensure that μ is a probability measure and the remaining m constraints correspond to requiring the measure μ has q_1, q_2, $\ldots$, q_m as its first m moments.

Our focus is where $\Omega = \{w_1, \ldots, w_n\} \subseteq \mathbb{R}$ is a finite set of real numbers and $\mathcal{B}$ is the power set of Ω. We assume that $\Omega = \{1, 2, \ldots, n\}$ and so $w_j = j$. In this setting, a measure μ can be represented by a nonnegative n-dimensional vector $(x_1, \ldots, x_n)$ where $\mu(w_j) = x_j$ and $f(w_j) = f_j$ for $j \in [1, n]$. We refer to the vector $(x_1, \ldots, x_n)$ as a *distribution* and often suppress the measure μ_x that it represents. We thus solve the *discrete moment problem* (DMP):

$$\text{(DMP)} \qquad \begin{aligned} &\max_{x \in \mathbb{R}^n} \sum_{j=1}^{n} f_j x_j && \text{(1a)} \\ &\text{s.t.} \sum_{j=1}^{n} w_j^i x_j = q_i \text{ for } i \in [0, m] && \text{(1b)} \\ &\mu_x \in \mathcal{P}. && \text{(1c)} \end{aligned}$$

Remark 1. In the results that follow, the choices of standard monomial moments w_j^i in (1b) (and not, e.g., binomial moments as in Prékopa (1990)) and the sample space $\Omega = \{1, 2, \ldots, n\}$ is largely taken for ease of notation. The structure of the moment constraints only affects the parameters of our optimal measures, not their general structure (which is detailed in Section 4). The main focus of this paper is uncovering the structure implicit in the choice of $\mathcal{P}$. The type of moments in (1b) are not of primary interest and, moreover, any choice of moments leaves (1b) linear in the x_j and thus amendable to our approach. Similarly, the choice of $\Omega = \{1, 2, \ldots, n\}$ (and not some other set of outcomes) does not affect the underlying structure of optimal measures implicit in $\mathcal{P}$, as long as the description of those measures are expressible in the elements of Ω. Indeed, some of the descriptions we consider for $\mathcal{P}$ (in particular, LC and IGFR distributions) are only defined for the case where $\Omega = \{1, 2, \ldots, n\}$. Hence, except when otherwise stated, one may think of $\Omega = \{1, 2, \ldots, n\}$ as the standard choice. See Remark 5 for additional discussion.

We study problem (1) under three specifications of the set of distributions $\mathcal{P}$ in constraint (1c).

Definition 1 (cf. Definition 2.2 in Canonne et al. (2018)). A distribution $x = (x_1, \ldots, x_n)$ is *discrete log-concave* (or simply *log-concave* or LC) if (i) for any $1 \leq k < j <$

$\ell \leq n$ such that $x_k x_\ell > 0$ then $x_j > 0$; and (ii) for all $j \in (1, n)$, $x_{j-1}x_{j+1} \leq x_j^2$. We let $\mathcal{P}_{\mathrm{LC}}$ denote the class of all LC distributions.

More precisely, (i) implies every LC distribution has a *consecutive support*; that is, there exists some $1 \leq k \leq \ell \leq n$ such that $x_j > 0$ for $j \in [k, \ell]$, and $x_j = 0$ otherwise. For an LC distribution x with support $[k, \ell]$, we must then ensure $x_{j-1}x_{j+1} \leq x_j^2$ holds for $j \in (k, \ell)$. At all other j, the inequality is trivial because at least one of x_{j-1} or x_{j+1} is zero.

Definition 2 (cf. Definition 2.4 in Canonne et al. (2018)). A distribution $x = (x_1, \ldots, x_n)$ has an *increasing failure rate* (IFR) if the failure-rate sequence $s_j := \frac{x_j}{\sum_{k=j}^n x_k}$ is a non-decreasing sequence; that is, $s_k \geq s_j$ for all $k \geq j$. We let $\mathcal{P}_{\mathrm{IFR}}$ denote the class of all IFR distributions.

Definition 3 (cf. Banciu and Mirchandani (2013)). A distribution $x = (x_1, \ldots, x_n)$ has an *increasing generalized failure rate* (IGFR) if the generalized-failure-rate sequence $g_j := j \frac{x_j}{\sum_{k=j}^n x_k} = j s_j$ is a non-decreasing sequence; that is, $g_k \geq g_j$ for all $k \geq j$. We let $\mathcal{P}_{\mathrm{IGFR}}$ denote the class of all IGFR distributions.

Remark 2. We remark that Canonne et al. (2018) develop approaches to test whether a given data set of samples of a discrete random variable statistically support that the underlying distribution is LC or IFR. These procedures can be naturally extended to test for the IGFR property. In other words, the LC, IFR, and IGFR properties are statistically verifiable. Further discussion of statistical tests is outside of the scope of this paper.

By definition, IFR and IGFR distributions always have consecutive support.

It is well known that $\mathcal{P}_{\mathrm{LC}}$ is a strict subset of $\mathcal{P}_{\mathrm{IFR}}$ (An 1997), and $\mathcal{P}_{\mathrm{IFR}}$ is a strict subset of $\mathcal{P}_{\mathrm{IGFR}}$ because the failure rate s_j is nonnegative. It is straightforward to see the sets $\mathcal{P}_{\mathrm{LC}}$, $\mathcal{P}_{\mathrm{IFR}}$ and $\mathcal{P}_{\mathrm{IGFR}}$ are nonconvex. However, they share one additional common feature that is critical to our approach.

Definition 4. A set R in $\mathbb{R}^n$ is *reverse convex* if $R = \mathbb{R}^n \setminus S$ for some convex set $S \subseteq \mathbb{R}^n$. A set R in $\mathbb{R}^n$ is said to be *reverse convex with respect to (w.r.t) a set* $T \subseteq \mathbb{R}^n$ if $R = T \setminus S$ for some convex set $S \subseteq \mathbb{R}^n$.

In the remainder of this section, we show that (DMP) with $\mathcal{P}$ equal to $\mathcal{P}_{\text{LC}}$, $\mathcal{P}_{\text{IFR}}$, or $\mathcal{P}_{\text{IGFR}}$ have reverse convex constraints w.r.t. $\mathbb{R}^n_+$. This common feature is leveraged to solve these related problems to global optimality in a unified framework. The seemingly straightforward generalization of reverse convexity to reverse convexity w.r.t. $\mathbb{R}^n_+$, however, can lead to significantly different analytical properties. For example, observe that if a function $f : \mathbb{R}^n \to \mathbb{R}$ is quasiconcave (over $\mathbb{R}^n$), its lower-level sets are reverse convex. However, a function whose lower-level sets are reverse convex w.r.t. some strict subset T of $\mathbb{R}^n$ need not be quasiconcave.

2.1. The moment problem over log-concave distributions

Consider problem (DMP) when $\mathcal{P} = \mathcal{P}_{\text{LC}}$. We separate the optimization over x into first determining a support (mapping to condition (i) of Definition 1) and then introducing inequalities of the form $x_{j-1}x_{j+1} \le x_j^2$ for j in that support (mapping to condition (ii) in Definition 1), thus yielding the two-stage optimization problem:

$$\begin{aligned}
&\max_{k,\ell:1\le k\le \ell\le n} \max_{x\in\mathbb{R}^n} \sum_{j=k}^{\ell} f_j x_j && \text{(2a)}\\
\text{(DMP-LC)}\quad &\text{s.t.} \sum_{j=k}^{\ell} w_j^i x_j = q_i \text{ for } i\in[0,m] && \text{(2b)}\\
&x_{j-1}x_{j+1} \le x_j^2 \quad \text{for } j \in (k,\ell) && \text{(2c)}\\
&x_j > 0 \text{ for } j\in[k,\ell] && \text{(2d)}\\
&x_j = 0 \text{ for } j\notin[k,\ell]. && \text{(2e)}
\end{aligned}$$

The strict constraints (2d) make the feasible region appear not to be closed. However, the following reformulation of (DMP-LC) reveals the feasible region can be described with non-strict inequalities involving polynomial functions of x (implying continuity) and are thus closed:

$$\begin{aligned}
&\max_{x\in\mathbb{R}^n} \sum_{j=1}^{n} f_j x_j && \text{(3a)}\\
\text{(DMP-LC')}\quad &\text{s.t.} \sum_{j=1}^{n} w_j^i x_j = q_i \text{ for } i\in[0,m] && \text{(3b)}
\end{aligned}$$

$$x_{j-u}^{v} x_{j+v}^{u} \le x_j^{u+v} \quad \text{for } j \in (1,n), u \in [1, j-1], v \in [1, n-j] \tag{3c}$$

$$x_j \ge 0 \text{ for } j \in [1,n]. \tag{3d}$$

Notice, moreover, that in (DMP-LC') there is no outer maximization over the support.

Proposition 1. *Problems (DMP-LC) and (DMP-LC') are equivalent.*

Proposition 2 below shows (DMP-LC') is a nonconvex optimization problem where constraint (3c) defines a reverse convex set w.r.t. $\mathbb{R}_+^n$.

Proposition 2. *The set* $\{(x,y,z) : x^u y^v > z^{u+v},\ x \ge 0, y \ge 0, z \ge 0\}$ *is convex for any positive integers* u *and* v.

Whereas the set $\{(x,y,z) : x^u y^v > z^{u+v},\ x \ge 0, y \ge 0, z \ge 0\}$ is convex, the set where nonnegativity is relaxed – that is, $S = \{(x,y,z) : x^u y^v > z^{u+v}\}$ – is *not* convex when u is an odd integer. Indeed, $(-2,-1,0)$ and $(1,2,0)$ are in S but $1/2(-2,-1,0) + 1/2(1,2,0) = (-1/2, 1/2, 0)$ is not in S, when u is odd.

2.2. The moment problem over IFR distributions

Consider problem (DMP) with $\mathcal{P} = \mathcal{P}_{\mathrm{IFR}}$. The following result illustrates a tight connection between the IFR case and the LC case. This result is known in the continuous case (see (Barlow and Proschan 1996, Chapter 2)). We provide details in the appendix for the discrete case.

Lemma 1. *A distribution* $x = (x_1, \ldots, x_n)$ *has an increasing failure rate if and only if its tail-probability sequence* $\{\bar{F}_1, \ldots, \bar{F}_n\}$ *is LC, where* $\bar{F}_j = \sum_{k=j}^{n} x_k$.

In the IFR case, (DMP) becomes

$$\max_{x \in \mathbb{R}^n} \sum_{j=1}^{n} f_j x_j \tag{4a}$$

(DMP-IFR)

$$\text{s.t.} \sum_{j=1}^{n} w_j^i x_j = q_i \text{ for } i \in [0,m] \tag{4b}$$

$$\frac{x_j}{\sum_{k=j}^{n} x_k} \text{ is non-decreasing in } j \tag{4c}$$

$$x_j \ge 0 \text{ for } j \in [1,n]. \tag{4d}$$

Here, we define 0/0 to be equal to 1 to allow for a finite support range. Using the transformation described in Lemma 1, where $y_j = \sum_{k=j}^n x_k$ denotes tail probabilities, we can reformulate (4) as

$$\text{(DMP-IFR')} \qquad \begin{aligned} \max_{y\in\mathbb{R}^n_+} \ & \sum_{j=1}^{n-1} f_j(y_j - y_{j+1}) + f_n y_n && \text{(5a)}\\ \text{s.t.} \ & \sum_{j=1}^{n} (w_j^i - w_{j-1}^i) y_j = q_i \quad \text{for } i\in[0,m] && \text{(5b)}\\ & y_{j-1}y_{j+1} \le y_j^2 \quad \text{for } j \in (1,n) && \text{(5c)}\\ & y_j - y_{j+1} \ge 0 \quad \text{for } j\in[1,n], && \text{(5d)} \end{aligned}$$

where w_0^i and y_{n+1} are set to 0. Constraint (5c) captures the log-concavity of the tail probabilities and (5d) captures the non-increasing property of tail probabilities, and the non-negativity property follows from the non-increasing property with $y_{n+1} = 0$. In the IFR case, no outer optimization over supports is necessary. The consecutiveness of supports is immediate from the monotonicity condition of the y_j. Indeed, once $y_j = 0$ for some j, $y_k = 0$ for all $k > j$ by monotonicity.

2.3. The moment problem over IGFR distributions

In the IGFR case, (DMP) with $\mathcal{P} = \mathcal{P}_{\text{IGFR}}$ becomes

$$\text{(DMP-IGFR)} \qquad \begin{aligned} \max_{x\in\mathbb{R}^n} \ & \sum_{j=1}^{n} f_j x_j && \text{(6a)}\\ \text{s.t.} \ & \sum_{j=1}^{n} w_j^i x_j = q_i \ \text{for } i\in[0,m] && \text{(6b)}\\ & j\frac{x_j}{\sum_{k=j}^n x_k} \ \text{is non-decreasing in } j && \text{(6c)}\\ & x_j \ge 0 \ \text{for } j\in[1,n]. && \text{(6d)} \end{aligned}$$

Similar to the IFR case, we use the transformation $y_j = \sum_{k=j}^n x_k$ to denote tail probabilities. The non-decreasing constraint $(j-1)\frac{y_{j-1}-y_j}{y_{j-1}} \le j\frac{y_j - y_{j+1}}{y_j}$ in (DMP-IGFR) is equivalent to $y_j y_{j-1} + (j-1)y_j^2 \ge j y_{j+1} y_{j-1}$. Observe that $y_j y_{j-1} = \left[\sqrt{j-1}\,y_j + \frac{y_{j-1}}{2\sqrt{j-1}}\right]^2 - (j-1)y_j^2 -$

$\frac{y_{j-1}^2}{4(j-1)}$ and $jy_{j+1}y_{j-1} = \left[j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}}\right]^2 - \left(j\sqrt{j-1}y_{j+1}\right)^2 - \frac{y_{j-1}^2}{4(j-1)}$. Combining these two identities with the inequality above yields that

$$\left(j\sqrt{j-1}y_{j+1}\right)^2 + \left[\sqrt{j-1}y_j + \frac{y_{j-1}}{2\sqrt{j-1}}\right]^2 \geq \left[j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}}\right]^2.$$

Because y_{j-1} and y_{j+1} are both nonnegative, this can be further reformulated as

$$\left\| \begin{pmatrix} j\sqrt{j-1}y_{j+1} \\ \sqrt{j-1}y_j + \frac{y_{j-1}}{2\sqrt{j-1}} \end{pmatrix} \right\| \geq j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}},$$

the reverse of the second-order cone (SOC) type constraint, as shown in Proposition 3 below.

PROPOSITION 3. *The set* $\left\{(x,y,z) : \left\| \begin{pmatrix} ax \\ by+cz \end{pmatrix} \right\| < ax + cz \right\}$ *is convex for any positive numbers* a*,* b*, and* c*.*

We can thus reformulate (6) as

$$\text{(DMP-IGFR')} \quad \begin{aligned} &\max_{y\in\mathbb{R}^n} \sum_{j=1}^{n-1} f_j(y_j - y_{j+1}) + f_n y_n && \text{(7a)} \\ &\text{s.t.} \sum_{j=1}^{n} (w_j^i - w_{j-1}^i) y_j = q_i \quad \text{for } i \in [0,m] && \text{(7b)} \\ &\left\| \begin{pmatrix} j\sqrt{j-1}y_{j+1} \\ \sqrt{j-1}y_j + \frac{y_{j-1}}{2\sqrt{j-1}} \end{pmatrix} \right\| \geq j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}} \quad \text{for } j \in (1,n) && \text{(7c)} \\ &y_j - y_{j+1} \geq 0 \quad \text{for } j \in [1,n], && \text{(7d)} \end{aligned}$$

where w_0^i, y_{n+1} are set to 0. Constraint (7c) formulates the defining property of IGFR distributions into a reverse SOC constraint, and (7d) captures the non-increasing and non-negativity property of tail probabilities (as y_{n+1} is set to 0).

3. Reverse convex optimization

In this section, we present a general class of problems (reverse convex optimization problems) that includes all the problems introduced in Section 2 as special cases. This class admits optimal extreme point solutions that are determined by setting a sufficient number of inequalities to equalities. This result is reminiscent of linear programming where extreme points have algebraic characterizations as basic feasible solutions.

Our analysis proceeds in two stages. First, we discuss a broad class of optimization problems that have optimal extreme point solutions. Second, we specialize this general class to a subclass of nonconvex optimization problems where the source of nonconvexity arises from reverse convex sets (see Definition 4).

3.1. Optimization over (nonconvex) compact sets

Let us first consider a general optimization problem:

$$\begin{aligned} &\min\ c(x) \\ &\text{s.t. } x \in S, \end{aligned} \tag{8}$$

where c is a lower semicontinuous and quasiconcave function and S is a nonempty, compact, but not necessarily convex, subset of $\mathbb{R}^n$. Note the results in this section can be generalized to any locally convex topological vector space in the sense of Aliprantis and Border (2006, Chapter 5). This generalization is not required for the study of the discrete moment problem but is potentially relevant for an exploration of the continuous case.

The goal of this subsection is to prove the following:

Theorem 1. *Problem (8) has an optimal extreme point.*

Recall that an *extreme point* of S is any point $x \in S$ where the set of d such that $x \pm \epsilon d \in S$ for some $\epsilon > 0$ is empty. Let $\operatorname{ext} S$ denote the extreme points of the set S. The special case to Theorem 1 where S is convex is well known and immediate from Aliprantis and Border (2006, Corollary 7.75):

Lemma 2. *If S is compact and convex then (8) has an optimal extreme point solution.*

The proof when S is not convex takes a couple more steps. The first step is to work with the closed convex hull $\overline{\operatorname{conv}}\, S$ of S, which is the intersection of all closed convex sets that contain S.

Lemma 3. *(Theorem 5.3 in Aliprantis and Border (2006)) The closed convex hull of a compact set is compact. In particular, $\overline{\operatorname{conv}}\, S$ is a compact convex set.*

The following lemma helps us leverage these results about closed convex hulls to learn about the original problem (8).

LEMMA 4. *Let S be a compact subset of $\mathbb{R}^n$. Then* $\operatorname{ext} \overline{conv} S \subseteq \operatorname{ext} S$.

We prove Theorem 1 using Lemmas 3 and 4 in Section EC.5 of the online supplement.

3.2. Reverse convex optimization problem with nonnegative constraints

The following lemma captures the essence of reverse convex optimization and serves as a visualization tool for understanding our main theoretical result below (see Theorem 2).

LEMMA 5. *Consider the optimization problem*

$$\begin{aligned} \min_{x\in\mathbb{R}^n}\ & c(x) \\ s.t.\ & x \in R_p,\ \textit{for } p \in [1,P], \end{aligned}$$

where c is a lower semicontinuous and quasiconcave function, $P \geq n$, and R_p are closed, reverse convex sets such that $F := \cap_p R_p$ is a nonempty and compact subset of $\mathbb{R}^n$. Then, there exists an optimal solution that lies on the boundary of at least n of the sets R_p.

Lemma 5 extracts some ideas from existing results (particularly from (Hillestad and Jacobsen 1980, Theorem 2)) and presents them in a clean, geometric form. To facilitate the understanding of this lemma, we further provide an intuitive graphical illustration in Figure 1.

Despite its elegance, this lemma is insufficient for our purposes. It only applies when the R_p are reverse convex. The argument breaks down if the R_p are reverse convex w.r.t. another convex set S, as needed for the problems in Section 2. In particular, when the convex set S is a polytope, even though we can use $R_p \cap S$ as a reverse convex set to replace R_p, it may contain the boundary of the polytope S. However, this boundary portion of S may not be the boundary of the feasible region $\cap_p R_p$ and thus may not contain an optimal solution as we might expect. The next discussion handles this idea elegantly in a special case.

A standard setting in reverse convex optimization is to consider a feasible region

$$F = \{x \in \mathbb{R}^n : f_p(x) \leq 0 \text{ for } p \in [1,P]\}$$

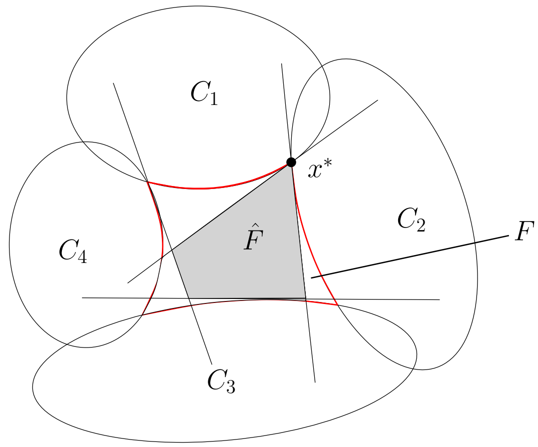

Figure 1 An illustration of the general reverse convex programming in $\mathbb{R}^2$. The feasible region F is the intersection of several R_p, where each R_p is the complement of a convex set C_p. After constructing the feasible polyhedron $\hat{F}$ (obtained via intersection of supporting hyperplanes of $\mathrm{cl}(C_p)$ that weakly separate x^*), we can show the optimal extreme point solution x^* lies on the boundary of at least two of the sets R_p, using the theory of basic feasible solutions in linear programming.

and assume properties on the functions f_p. These properties typically include differentiability assumptions (so that gradients are defined) and some form of concavity (the weakest being quasiconcavity) over the whole space $\mathbb{R}^n$ (see e.g., Theorem 2 in Hillestad and Jacobsen (1980)). Under these concavity assumptions, the lower-level sets of f_p are reverse convex (with respect to the whole space), and Lemma 5 applies so that extreme points are determined by a minimum number of tight constraints of the form $f_p(x) = 0$. Unfortunately, those results do not directly apply in our setting, because our functions are not quasiconcave over the whole domain. Instead, our problems involve functions whose lower-level sets are reverse convex *w.r.t. the nonnegative orthant.*

These considerations motivate us to establish a more general theory of reverse convex optimization. In particular, we analyze the following problem:

$$\begin{aligned} \min\ & c(x) \\ \text{s.t. } & Ax = b \\ & f_p(x) \le 0 \text{ for } p \in [1, P] \end{aligned} \tag{Rev-Cvx}$$

$$x \geq 0,$$

where c and the f_p are functions from $\mathbb{R}^n$ to $\mathbb{R}$, A is an m by n matrix, and for $1 \leq p \leq P$, the set $\{x : f_p(x) \leq 0\}$ is reverse convex w.r.t. the nonnegative orthant $\mathbb{R}^n_+$.

Assumption 1. *We make the following additional technical assumptions on* (Rev-Cvx)*:*

(i) the objective function $c(x)$ is continuous and quasiconcave,

(ii) the matrix A is full-row rank with $n \geq m$,

(iii) for each p, f_p is differentiable, and

(iv) the feasible region $F = \{x \in \mathbb{R}^n_+ : Ax = b,\ f_p(x) \leq 0,\ p = 1, \ldots, P\}$ is nonempty and compact.

Under these assumptions, we get our desired result.

Theorem 2. *Consider an instance of* (Rev-Cvx) *where Assumption 1 holds. Then, an optimal extreme point solution exists. Moreover, for any extreme point optimal solution x^*, at least $n - m$ of the following $P + n$ inequalities are tight:*

$$f_p(x^*) \leq 0 \text{ for } p \in [1, P] \tag{9}$$

$$x^*_j \geq 0 \text{ for } j \in [1, n]. \tag{10}$$

The proof of theorem largely follows the geometric intuition captured in Figure 1. At its core, it involves defining separating hyperplanes and inscribing a polyhedral set $\hat{F}$ inside the feasible region. Then, the equivalence of extreme points and basic feasible solutions for the polyhedron $\hat{F}$ is leveraged to establish the result.

However, the proof has additional technical challenges. It must make sense of how inequalities that describe the orthant $\mathbb{R}^n_+$ interact with the gradients of the constraint functions f_p. Moreover, the affine equality constraints, $Ax = b$, that correspond to the moment conditions in (1), force us to work within the affine space defined by these constraints for much of the proof. Technical details of the proof are found in the electronic companion.

Remark 3. One might wonder why the nonnegativity constraints in (Rev-Cvx) are not included as reverse convex constraints directly and denote them accordingly by f_p. We do this for two reasons. One is to emphasize that f_p are only reverse convex with respect to the positive orthant, which is highlighted by including the constraints $x_j \geq 0$ explicitly. The second reason is that in later arguments, we will show that in the appropriate formulation of our moment problems we can argue the nonnegativity constraints are not tight at an optimal solution (see, for instance, the comment following (12)). The fact that these constraints are not tight allows us to say that at least $n-m$ of the constraints $f_p(x^*) \leq 0$ are binding at an optimal extreme point x^*. Setting these constraints to be tight yields the desired structure on our optimal solutions.

4. Characterizing optimal extreme point solutions in the discrete moment problem

Theorems 1 and 2 are powerful tools for analyzing the moment problems we discussed in Section 2. They allow us to characterize the structure of optimal extreme point solutions. In the following three subsections, we analyze the LC, IFR, and IGFR distributions cases from Sections 2.1, 2.2 and 4.3, respectively.

Our analysis has the following general pattern. Each problem has two alternate formulations, with one indicated by a "prime". In the LC case, these two formulations are (DMP-LC) and (DMP-LC'). The "prime" formulation has a closed and compact feasible region that allows us to leverage Theorem 1 to show the existence of an optimal extreme point solution x^*. With x^* in hand, we apply Theorem 2 to a small adjustment of the "non-prime" formulation that replaces strict inequalities with non-strict inequalities based on the support of x^*. Theorem 2 implies that a certain number of constraints are tight, including some number of the reverse convex constraints (e.g., (2c) in (DMP-LC)). Making these constraints tight determines the structure of the optimal extreme point solutions. Although the proofs follow a similar pattern, we give details in the appendix for each case. There are some important details in each case that need special care, so we provide careful details for completeness.

4.1. Log-concavity

Recall the two alternate formulations, (DMP-LC) and (DMP-LC'). In particular, recall that there are $m+1$ moment constraints in (2b) and (3b).

Theorem 3. *Every feasible instance of (DMP-LC) has an optimal extreme point solution. Moreover, when $n \geq m$, every optimal extreme point solution x^* has the following structure: there exist (i) integers u_i and v_i for $i \in [1,m]$ with $k = u_1 < v_1 = u_2 < v_2 \cdots < v_{m-1} = u_m < v_m = \ell$ where $[k,\ell]$ is the support of x^* and (ii) real parameters $\alpha_i > 0$, $r_i > 0$ with $\alpha_{i+1} = \alpha_i r_i^{v_i - u_i}$ for $i \in [1,m]$ such that*

$$x_j^* = \begin{cases} \alpha_i r_i^{j-u_i} & \text{for } j \in [u_i, v_i] \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

That is, an optimal solution to (DMP-LC) exists with piecewise geometric structure with (at most) m pieces.

Proof. Consider the (DMP-LC') representation of the problem. The 0-th order moment constraint ((3b) for $i=0$) is $\sum_{i=1}^n x_i = 1$, which, along with the nonnegative constraints (3d), implies the feasible region of the problem (DMP-LC') is bounded. To leverage Theorem 1 to show that an optimal extreme point solution to (DMP-LC') exists, it thus suffices to show the feasible region is closed. This follows since all of the constraint functions are continuous in the decision vector x. By Proposition 1, (DMP-LC) also has an optimal extreme point.

Let x^* be any extreme optimal solution and, for simplicity, we assume its support is $[1,n]$ (the general case of supposing $[k,\ell]$ with $1 < k < \ell < n$ follows analogously). Note that when $n \leq m$, there are at most m points in the interval $[1,n]$, where each point x_j, $j \in [1,n]$ could be viewed as a single piece, and the conclusion readily follows. Therefore, in the remainder of the proof, we assume $n \geq m+1$.

Let $\underline{x} := \min\{x_j^* : j \in [1,n]\} > 0$ and define the following problem:

$$\max_{x \in \mathbb{R}^n} \sum_{j=1}^n f_j x_j \tag{12a}$$

$$\text{s.t.} \sum_{j=1}^n w_j^i x_j = q_i \text{ for } i \in [0,m] \tag{12b}$$

$$x_{j-1}x_{j+1} \leq x_j^2 \quad \text{for } j \in (1,n) \tag{12c}$$

$$x_j \geq \underline{x}/2 \text{ for } j \in [1,n]. \tag{12d}$$

$$x_j \geq 0 \text{ for } j \in [1,n] \tag{12e}$$

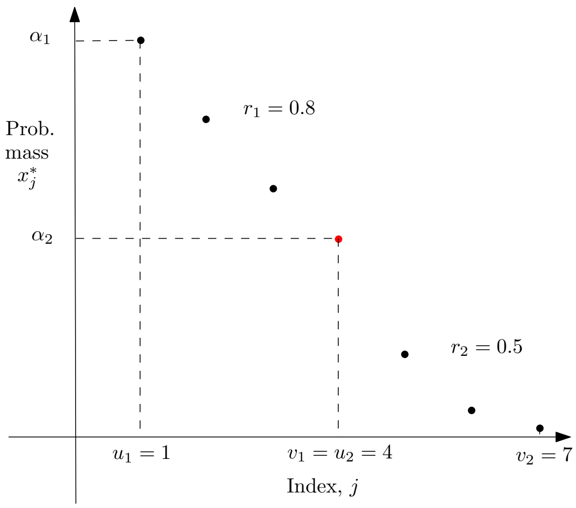

Figure 2 Piecewise geometric structure of optimal extreme point solutions for a problem with $m=2$.

Note that (12) is a restriction of (DMP-LC) with a given support and replacing the strict inequalities in (2d) with non-strict inequalities in (12d). We have also added redundant nonnegativity constraints in (12e) so that it fits our reverse convex programming framework as in Theorem 2. Note also that x^* is an extreme optimal solution to (DMP-LC) and it is feasible to (12); hence, x^* is an extreme optimal solution to (12).

At the extreme point x^*, the constraints (12d) and (12e) cannot be tight, so the application of Theorem 2 implies that at most $m-1$ of the (12c) constraints are not tight at x^*. These non-tight indexes can divide the interval $[1,n]$ into at most m pieces, and within each piece, we have $x^*_{j-1}x^*_{j+1} = (x^*_j)^2$, for $j \in [u_i, v_i]$, where u_i and v_i are the left and right endpoints of piece i of the domain. Note that such a system implies $x^*_j = r_i^{j-u_i} x^*_{u_i}$ for $j \in [u_i, v_i]$. Setting $\alpha_i = x^*_{u_i}$ yields the form (11). $\square$

The piecewise geometric form (11) of optimal extreme point distributions to (DMP-LC) is illustrated in Figure 2.

Remark 4. Note that a basic count of tight constraints is able to deliver the piecewise geometric structure since the number of constraints in (DMP-LC) for a given support is small compared to the number of variables. Consider support $[1,n]$ in (DMP-LC). Theorem 2 implies $n-m$ of the $2n-2$ constraints in (2c)–(2d) are tight. Since all constraints in (2d) are strict (shown carefully in the proof), this implies all $n-m$ tight constraints are from (2c) and thus of the form $x_{j-1}x_{j+1} \le x_j^2$. Setting $n-m$ of these constraints to equality directly yields the geometric structure (11).

REMARK 5. This remark builds on Remark 1 on the choice of the structure of the moment constraints in (1b), in the specific context of this section. Notice the main contribution of Theorem 3 is how to handle the shape constraints (2c). We use Theorem 2 to argue a sufficient number of the shape constraints are *tight*, lending the optimal solution structure, as seen in (11). Note this structure holds independently of the form of the moments in constraint (2b). A different choice other than the power moments will only impact the parameters in (11) (i.e., the specific α_i and r_i) and not the overall structure. In our numerical experiments, we focus on computing parameters given the power moments in (1b) for their simplicity; however, these methods could be adapted for more general moments, possibly with additional effort needed for computation.

We do remark that changing the domain from $\Omega = \{1, 2, \ldots, n\}$ to some other domain could present issues, largely because we do not have definitions of LC and IGFR in these cases. Indeed, all reference to discrete LC distributions we could find in the literature (including Canonne et al. (2018), Johnson and Goldschmidt (2006), Wellner (2013), Johnson et al. (2013), Balabdaoui et al. (2011), Saumard and Wellner (2014)) take $\Omega = \{1, 2, \ldots, n\}$. That domain is hard-coded into the definition of g_j in Definition 3.

4.2. Increasing failure rate

Recall the formulation (DMP-IFR') of the IFR moment problem in Section 2.2 with $y_j = \sum_{k=j}^n x_k$. We will show the optimal solution has a similar structure to the LC case, again using Theorems 1 and 2.

Here we notice two facts. First, by the IFR constraint and the non-increasing property of y_j, any feasible solution y has consecutive support naturally, and the support starts from $y_1 = 1$. This case is different from the LC case. Second, if there is some ℓ such that $y_\ell = y_{\ell+1} > 0$, this scenario combined with the constraint $y_{\ell-1}y_{\ell+1} \le y_\ell^2$ indicates we have $y_{\ell-1} \le y_\ell$. However, we also have $y_{\ell-1} \ge y_\ell$ in the problem's constraints, which means $y_\ell = y_{\ell+1} > 0$ implies $y_{\ell-1} = y_\ell$. Then, by induction, we have $y_1 = \cdots = y_{\ell+1} = 1$.

Combining the two facts above, the interval $[1,n]$ can be divided into three consecutive parts: $[1,k),[k,\ell),[\ell,n]$, where we have $y_1=\cdots=y_k=1,\ 1=y_k>\cdots>y_\ell=0,\ 0=y_\ell=\cdots=y_n$; that is, an all-one interval, a strictly decreasing interval, and an all-zero interval. Further, the optimal solution in the middle interval has a more detailed characterization stated here.

THEOREM 4. *Every feasible instance of (DMP-IFR') has an optimal extreme point solution. Moreover, when $n\ge m+1$, for every optimal extreme point solution y^*, there exist integers $1\le k\le \ell$ such that $y_j=1$ when $j\le k$, $y_j=0$ when $j\ge \ell$. The interval $[k,\ell)$ can be divided as follows. There exist (i) integers u_i and v_i for $i\in[1,m]$ with $k=u_1<v_1=u_2<v_2\cdots<v_{m-1}=u_m<v_m=\ell-1$ (ii) real parameters $\alpha_i>0,\ 0<r_i<1$ with $\alpha_{i+1}=\alpha_i r_i^{v_i-u_i}$ for $i\in[1,m]$ such that*

$$y_j^*=\begin{cases}\alpha_i r_i^{j-u_i} & \text{for } j\in[u_i,v_i]\\ 1 & j\le k\\ 0 & j\ge \ell.\end{cases}\tag{13}$$

4.3. Increasing generalized failure rate

Similar to the IFR case, we establish a similar structural result for the IGFR case. The proof follows the pattern described at the outset of this section.

THEOREM 5. *Every feasible instance of (DMP-IGFR') has an optimal extreme point solution. Moreover, when $n\ge m+1$, for every optimal extreme point solution y^*, there exist integers $1\le k\le \ell$ such that $y_j=1$ when $j\le k$, $y_j=0$ when $j\ge \ell$. The interval $[k,\ell)$ can be divided as follows. There exist (i) integers u_i and v_i for $i\in[1,m]$ with $k=u_1<v_1=u_2<v_2\cdots<v_{m-1}=u_m<v_m=\ell-1$ (ii) real parameters $\alpha_i>0,\ 0<r_i<1$ with $\alpha_{i+1}=\alpha_i\prod_{k=1}^{v_i-u_i}(1-\frac{r_i}{k+u_i-1})$ for $i\in[1,m]$ such that*

$$y_j^*=\begin{cases}\alpha_i & \text{for } j=u_i\\ \alpha_i\prod_{k=1}^{j-u_i}(1-\frac{r_i}{k+u_i-1}) & \text{for } j\in(u_i,v_i]\\ 1 & j\le k\\ 0 & j\ge \ell.\end{cases}\tag{14}$$

Observe that the tail probabilities are no longer piecewise geometric, as in previous cases, but nonetheless have an attractive structure that we can work with in later sections.

5. Numerical results

In this section, we use the results in Section 4 to numerically solve a representative sample of moment problems and a robust newsvendor problem. We first focus on a practical method to solve the moment problem over LC distributions, IFR distributions, or IGFR distributions with *two moments* as a proof of concept of our approach. Then, we provide some preliminary numerical results of our method on a binomial moment-constrained problem and a robust newsvendor problem.

5.1. Computational approach for the moment problems

Our computational approach is based on the structure of optimal extreme point solutions in Theorems 3 to 5. According to those theorems, when restricting attention to the problem with two moment constraints, an *optimal two-piece* distribution exists for the corresponding moment-constrained problem, (DMP-LC), (DMP-IFR), or (DMP-IGFR). According to (11), (13), and (14), we can restrict the search on the seven-dimensional decision space– k, v_1, ℓ, α_1, α_2, r_1, and r_2– to construct a solution that satisfies the constraints of the problem with the largest objective value. The transformed objective function over this seven-dimensional decision space is denoted $\hat{f}(\cdot)$. A more traditional approach to solving (11), (13), and (14) would take x or y as the decision variable and solve them directly. The resulting problems are nonconvex and (potentially) high dimensional if n is large, whereas our approach remains low dimensional as n grows.

We first implement a normalization step to simplify the analysis. Recall that an instance of a two-moment problem is specified by the elements of the sample space $\Omega = \{w_1, \dots, w_n\}$ and the moments q_1 and q_2. In the rest of this section, we shift and scale the elements of the sample space so that the resulting distribution has mean $q_1' = 0$ and variance $q_2' = 1$. For each $j \in [1, n]$, this means subtracting the mean q_1 from w_j and scaling the result by $\epsilon := 1/\sqrt{q_2 - q_1^2}$. The resulting sample space is $\Omega' = \{w_1', \dots, w_n'\} = \{w_1\epsilon - q_1\epsilon, w_2\epsilon - q_1\epsilon, \dots, w_n\epsilon - q_1\epsilon\}$. That is, $w_j' = w_j\epsilon - q_1\epsilon$ for $j \in [1, n]$. Again, for simplicity we assume, as in Section 2, that $w_j = j$ so that we have $w_j' = j\epsilon - q_1\epsilon$.

In the aforementioned seven-dimensional decision space, the first three variables concern the domain: k and ℓ describe the support, and v_1 describes the "break point" between the two pieces of the underlying distribution. For fixed k, ℓ, and v_1, we note $\alpha_2 = \alpha_1 r_1^{v_1-k}$ for the log-concave and IFR distributions, whereas $\alpha_2 = \alpha_1 \prod_{w=1}^{v_1-k}(1 - \frac{r_1}{w+k-1})$ for the IGFR distribution. Furthermore, note that $\alpha_1 = y_k = 1$ for the IFR and IGFR distributions, whereas for the LC distribution the 0-th moment condition of

$$\sum_{j=k}^{v_1-1} \alpha_1 r_1^{j-k} + \sum_{j=v_1}^{\ell} \alpha_1 r_1^{v_1-k} r_2^{j-v_1} = 1 \tag{15}$$

implies $\alpha_1 = 1/(\sum_{j=k}^{v_1-1} r_1^{j-k} + \sum_{j=v_1}^{\ell} r_1^{v_1-k} r_2^{j-v_1})$. Therefore, we can represent α_1 and α_2 explicitly in terms of k and ℓ and v_1. Consequently the first- and second-order moment constraints give us two equations, $g_1(r_1, r_2) = 0$ and $g_2(r_1, r_2) = 0$, in two unknowns r_1 and r_2. In fact, the equation $g_1(r_1, r_2) = 0$ implies the existence of a function $h(\cdot)$ such that $r_2 = h(r_1)$. Replacing r_2 with $h(r_1)$ in $g_2(\cdot)$ yields a new equation $g_2(r_1, h(r_1)) = 0$ with one unknown r_1. This equation can be solved by any solution method for a one-dimensional nonlinear equation. In this paper, we adopt Newton's method. Unfortunately, we are unable to get analytical forms of $h(r_1)$ and $g(r_1, h(r_1))$. However, we can evaluate the values of $h(r_1)$, $g_2(r_1, h(r_1))$, and $\frac{\mathrm{d}g_2(r_1, h(r_1))}{\mathrm{d}r_1}$ for any r_1, which is sufficient for applying Newton's method. The treatments for computing these values is different for each of the three cases (LC, IFR, and IGFR) and are explained separately below. Our method to solve the two moment problem is made explicit in Algorithm 1.

Algorithm 1 has $O(n^3)$ iteration complexity. Giving run-time complexity is difficult because we apply Newton's method in each iteration to solve a one-dimensional nonlinear equation, and Newton's method does not have a global convergence rate result. However, we provide the total realized running times of our algorithms for specific instances in Figure 3 to justify the efficiency of our approach.

Our main focus is problems with two moments. Additional details on this case for LC, IFR, and IGFR problems are provided in the next subsections. In principle, however, our approach can be generalized to problems with m moments. We provide a brief discussion here for the LC case (the IFR or IGFR shape-constrained cases follow from similar arguments). According to Theorem 3, at most m geometric pieces exist and we can enumerate the endpoints v_i and u_i of each piece. The probability mass α_i at the starting point and

Algorithm 1 Solving the two-moment problem

Initialization: $f^* = -\infty$;
for $1 \le k \le v_1 \le \ell \le n$ [enumerate all possible k, l and v_1] **do**
  Solve equation $g_2(r_1, h(r_1)) = 0$ by Newton's method;
  if a solution r_1 exists **then**
    Compute $r_2 = h(r_1)$;
    Let $\alpha_1 = 1$ for IFR/IGFR ($\alpha_1 = 1/(\sum_{j=k}^{v_1-1} r_1^{j-k} + \sum_{j=v_1}^{\ell} r_1^{v_1-k} r_2^{j-v_1})$ for LC);
    Compute $\alpha_2 = \alpha_1 r_1^{v_1-k}$ for LC/IFR ($\alpha_2 = \alpha_1 \prod_{w=1}^{v_1-k}(1 - \frac{r_1}{w+k-1})$ for IGFR);
    if $k, v_1, \ell, \alpha_1, \alpha_2, r_1, r_2$ satisfy the LC/IFR/IGFR constraint at v_1 **then**
      Compute the objective value $f^{Temp} = \hat{f}(k, v_1, \ell, \alpha_1, \alpha_2, r_1, r_2)$;
      if $f^{Temp} > f^*$ **then**
        Update $f^* = f^{Temp}$ and $(k^*, v_1^*, \ell^*, \alpha_1^*, \alpha_2^*, r_1^*, r_2^*) = (k, v_1, \ell, \alpha_1, \alpha_2, r_1, r_2)$.
      end if
    end if
  end if
end for
Return optimal value f^* and optimal solution $(k^*, v_1^*, \ell^*, \alpha_1^*, \alpha_2^*, r_1^*, r_2^*)$.

the slope r_i of each piece remains to be found. In fact, it suffices to treat α_1 and r_i's as independent variables because the other α_i are determined by the relation $\alpha_{i+1} = \alpha_i r_i^{v_i - u_i}$. Moreover, the choice of α_1 is determined (given the $r_1, \ldots, r_m$) by the 0-th moment condition in (2b). Consequently, given the endpoints v_i and u_i, we solve m nonlinear equations introduced by the m moment constraints in m unknowns $r_1, \cdots, r_m$. Algorithm 1 specifies a way to solve these equations when $m = 2$. For general m, we can resort to algorithms for solving general nonlinear equations. In particular, we can transform the nonlinear-equation problem to a nonlinear least-squares problem and then apply the Gauss-Newton method or Levenberg-Marquardt methods to find numerical solutions. Finding additional structure that can speed up these computations may be possible, but analysis in this direction is beyond our scope.

5.1.1. Log-concave distribution Now, we explain how to solve the equation $g_2(r_1,h(r_1))=0$ in Algorithm 1 when the underlying distribution is LC. To implement Newton's method, we need to evaluate $h(r_1)$, $g_2(r_1,h(r_1))$ and $\frac{\mathrm{d}g_2(r_1,h(r_1))}{\mathrm{d}r_1}$ numerically. In this case, we are given the support $[k,\ell]$ and break point v_1 in Algorithm 1. For ease of computation, we perform a variable transformation by letting $\rho=\alpha_2=\alpha_1 r_1^{v_1-k}$, $r_2=e^{\alpha}$, and $r_1=e^{-\beta}$ for nonnegative scalars α and β such that the 0-th, first and second moment conditions (recall that via normalization, $q_1'=0$, $q_2'=1$ and sample point $w_j'=j\epsilon-q_1\epsilon$) amount to

$$\sum_{j=1}^{\tilde{k}}\rho e^{\beta j}+\sum_{j=1}^{\tilde{\ell}}\rho e^{\alpha j}+\rho=1 \tag{16}$$

$$\sum_{j=1}^{\tilde{k}}\rho e^{\beta j}(a-j\epsilon)+\sum_{j=1}^{\tilde{\ell}}\rho e^{\alpha j}(a+j\epsilon)+\rho a=0. \tag{17}$$

$$\sum_{j=1}^{\tilde{k}}\rho e^{\beta j}(a-j\epsilon)^2+\sum_{j=1}^{\tilde{\ell}}\rho e^{\alpha j}(a+j\epsilon)^2+\rho a^2=1, \tag{18}$$

where $a=w_{v_1}'=v_1\epsilon-q_1\epsilon$. We re-index the sums and set $\tilde{k}=v_1-k$ and $\tilde{\ell}=\ell-v_1$. By first eliminating ρ, we get two equations in two unknowns:

$$g_1(\alpha,\beta):=\sum_{j=1}^{\tilde{k}}e^{\beta j}(a-j\epsilon)+\sum_{j=1}^{\tilde{\ell}}e^{\alpha j}(a+j\epsilon)+a=0 \tag{19}$$

$$g_2(\alpha,\beta):=\sum_{j=1}^{\tilde{k}}e^{\beta j}+\sum_{j=1}^{\tilde{\ell}}e^{\alpha j}+1-\sum_{j=1}^{\tilde{k}}e^{\beta j}(a-j\epsilon)^2-\sum_{j=1}^{\tilde{\ell}}e^{\alpha j}(a+j\epsilon)^2-a^2=0. \tag{20}$$

Next, we want to show that, given an α, a unique choice of β exists such that $g_1(\alpha,\beta)=0$, which further implies the existence of the function $h(\cdot)$ such that $\beta=h(\beta)$. We achieve this task by exploring monotonicity properties of g_1. First, a direct computation yields

$$\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}=\sum_{j=1}^{\tilde{\ell}}j\cdot(a+j\epsilon)e^{\alpha j}=\mathbb{E}\left[\frac{X}{\rho}\cdot\frac{X-a}{\epsilon}\cdot\mathbb{1}_{X>a}\right] \tag{21}$$

$$\frac{\partial g_1(\alpha,\beta)}{\partial\beta}=\sum_{j=1}^{\tilde{k}}j\cdot(a-j\epsilon)e^{\beta j}=\mathbb{E}\left[\frac{X}{\rho}\cdot\frac{a-X}{\epsilon}\cdot\mathbb{1}_{X<a}\right], \tag{22}$$

where X is the discrete random variable with distribution $x=(x_1,\ldots,x_n)$. We then use the following technical lemma.

LEMMA 6. *Suppose the polynomial $\phi(z) = \sum_{j=1}^{M} a_j z^{i_j}$ with $z \in \mathbb{R}$ satisfies*

$$a_1 \le a_2 \le \cdots \le a_M \quad \text{and} \quad 1 \le i_1 \le i_2 \le \cdots \le i_M, \tag{23}$$

where the a_i's are not all zero. Then, $\phi(z)$ has at most one root when $z > 0$ and is increasing on $\{z \mid \phi(z) \ge 0\}$.

It follows from (17) that $\mathbb{E}[X \cdot \mathbb{1}_{X>a}] = \sum_{j=1}^{\tilde{k}} e^{\alpha j}(a + j\epsilon) \ge 0$ and $\mathbb{E}[X \cdot \mathbb{1}_{X<a}] = \sum_{j=1}^{\tilde{\ell}} e^{\beta j}(a - j\epsilon) \le 0$. Now, apply Lemma 6 to the polynomials $\sum_{j=1}^{\tilde{k}} e^{\alpha j}(a + j\epsilon)$ and $-\sum_{j=1}^{\tilde{\ell}} e^{\beta j}(a - j\epsilon)$, respectively (in the former, z is e^{α} and $a_j = a + j\epsilon$). Supposing roots exist to these polynomials, define

$$\alpha_0 = \min\{\alpha : \sum_{j=1}^{\tilde{\ell}} e^{\alpha j}(a + j\epsilon) = 0\} \quad \text{and} \quad \beta_0 = \min\{\beta : \sum_{j=1}^{\tilde{k}} e^{\beta j}(a - j\epsilon) = 0\}$$

such that $\mathbb{E}[X \cdot \mathbb{1}_{X>a}] \ge 0$ if and only if $\alpha \ge \alpha_0$, and $\mathbb{E}[X \cdot \mathbb{1}_{X<a}] \le 0$ if and only if $\beta \ge \beta_0$. If roots do not exist, set $\alpha_0 = -\infty$ and/or $\beta_0 = -\infty$. Thus, it suffices to focus on the region where $\alpha \ge \alpha_0$ and $\beta \ge \beta_0$. As a result, when $\alpha \ge \alpha_0$,

$$\mathbb{E}[X \cdot (X - a)\mathbb{1}_{X>a}] = \begin{cases} \mathbb{E}[(X-a)^2 \mathbb{1}_{X>a}] + \mathbb{E}[a \cdot (X-a)\mathbb{1}_{X>a}] \ge 0, & \text{if } a \ge 0 \\ \mathbb{E}[X^2 \mathbb{1}_{X>a}] - a\,\mathbb{E}[X \cdot \mathbb{1}_{X>a}] \ge 0, & \text{if } a < 0. \end{cases}$$

Similarly for $\beta \ge \beta_0$, we have

$$\mathbb{E}[X \cdot (a - X)\mathbb{1}_{X<a}] = \begin{cases} \mathbb{E}[-X^2 \mathbb{1}_{X<a}] + a\,\mathbb{E}[X \cdot \mathbb{1}_{X<a}] \le 0, & \text{if } a \ge 0 \\ \mathbb{E}[-(X-a)^2 \mathbb{1}_{X<a}] + \mathbb{E}[a \cdot (a-X)\mathbb{1}_{X<a}] \le 0, & \text{if } a < 0. \end{cases} \tag{24}$$

In summary, we have $\frac{\partial g_1(\alpha,\beta)}{\partial \alpha} \ge 0$ when $\alpha \ge \alpha_0$, and $\frac{\partial g_1(\alpha,\beta)}{\partial \beta} \le 0$ when $\beta \ge \beta_0$. This monotonicity yields our desired property that we can identify a mapping $h(\cdot)$ such that $\beta = h(\alpha)$. Plugging this identity into (20) yields the equation $g_2(\alpha, h(\alpha)) = 0$, which corresponds to equation $g_2(r_1, h(r_1)) = 0$ in Algorithm 1. To apply Newton's method, we must find the derivative with respect to α. Observe that

$$\frac{\partial g_2(\alpha,\beta)}{\partial \alpha} = \sum_{j=1}^{\tilde{\ell}} j \cdot e^{\alpha j} - \sum_{j=1}^{\tilde{\ell}} j \cdot (a + j\epsilon)^2 e^{\alpha j} = \mathbb{E}\left[\frac{1 - X^2}{\rho} \cdot \frac{X - a}{\epsilon} \cdot \mathbb{1}_{X>a}\right] \tag{25}$$

and

$$\frac{\partial g_2(\alpha,\beta)}{\partial \beta} = \sum_{j=1}^{\tilde{k}} j \cdot e^{\beta j} - \sum_{j=1}^{\tilde{k}} j \cdot e^{\beta j}(a - j\epsilon)^2 = \mathbb{E}\left[\frac{1 - X^2}{\rho} \cdot \frac{a - X}{\epsilon} \cdot \mathbb{1}_{X<a}\right], \tag{26}$$

and so

$$\frac{\mathrm{d}g_2(\alpha,h(\alpha))}{\mathrm{d}\alpha}=\frac{\partial g_2(\alpha,\beta)}{\partial\alpha}+\frac{\partial g_2(\alpha,\beta)}{\partial\beta}\cdot\frac{\partial h(\alpha)}{\partial\alpha}=\frac{\partial g_2(\alpha,\beta)}{\partial\alpha}+\frac{\partial g_2(\alpha,\beta)}{\partial\beta}\cdot\frac{-\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}}{\frac{\partial g_1(\alpha,\beta)}{\partial\beta}}. \quad (27)$$

Using this derivative, Newton's method on the interval $[\alpha_0,+\infty]$ of real numbers finds all roots of $g_2(\alpha,h(\alpha))$.

It is well known that Newton's method has local quadratic convergence, provided that the gradient of function is Lipschitz continuous (see Chapter 2 in Dennis Jr and Schnabel (1996)). Unfortunately, we are not able to prove that the gradient $\frac{d}{d\alpha}g_2(\alpha,h(\alpha))$ in our problem is Lipschitz continuous. However, we can establish that the gradient is continuous. This requires part the following technical lemma.

Lemma 7. *Suppose the discrete random variable X has diameter D, i.e., $|X|\le D$ for all X. When the distribution of X has positive mass on both sides of a then the absolute value of $\frac{d}{d\beta}g_1(\alpha,\beta)$ is uniformly bounded below.*

Remark 6. If the distribution of X is only on one side of a (potentially including a) it must be single piece log-concave distributed, where the densities are determined by (19) and Lemma 6.

Theorem 6. *When the distribution of X has positive mass on both sides of a, the gradient $\frac{d}{d\alpha}g_2(\alpha,h(\alpha))$ is continuous.*

It is easy to see why $\frac{d}{d\alpha}g_2(\alpha,h(\alpha))$ is continuous. Given the expression of $\frac{d}{d\alpha}g_2(\alpha,h(\alpha))$ in (27), this follows from the continuity of $\frac{\partial g_2(\alpha,\beta)}{\partial\alpha}$, $\frac{\partial g_2(\alpha,\beta)}{\partial\beta}$, $\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}$, $\frac{\partial g_1(\alpha,\beta)}{\partial\beta}$, and the fact that $|\frac{\partial g_1(\alpha,\beta)}{\partial\beta}|$ has a uniform lower bound.

The fact that $\frac{d}{d\alpha}g_2(\alpha,h(\alpha))$ is continuous (and thus $g_2(\alpha,h(\alpha))$ is continuously differentiable in α) is useful for showing convergence of Newton's method, as shown in the following result.

Theorem 7. *Suppose the function φ is continuously differentiable and γ is a simple root of ϕ (i.e. $\varphi(\gamma)=0$ and $\varphi'(\gamma)\neq 0$). Then there exists a $\delta>0$ such that $|z_0-\gamma|\le\delta$ implies that the sequence $\{z_k\}$ generated by the Newton's method applied to φ satisfies the following two properties: (i) $|z_{k+1}-\gamma|<|z_k-\gamma|\le\delta$ and (ii) $\{z_k\}$ converges superlinearly.*

Here ϕ plays the role of $g_2(\alpha, h(\alpha))$ and the target root for α is denoted γ. Since the roots we seek are of the system of two polynomials (the system (19)–(20)) the target root γ is almost surely simple. Discussing the degenerate case of Newton's method on a non-simple root is beyond the scope of our paper.

Newton's method starting from a given initial point could converge to one root of $g_2(\alpha, h(\alpha))$, which is not necessarily an optimal solution to the original moment problem. Therefore, we need to run Newton's with multiple initial points to find all roots and pick the best one. However, the question remains as to how many initial points are required to get an approximate root within any given tolerance level $\hat{\epsilon}$. In the theorem below, we shall show $|\frac{d}{d\alpha} g_2(\alpha, h(\alpha))|$ is uniformly upper bounded by some constant L. Consequently, when the current solution x_k is not an approximate root (i.e. $|g_2(x_k, h(x_k))| > \hat{\epsilon}$), the Newton step size is $|-\frac{g_2(x_k, h(x_k))}{g_2' 2(\alpha, h(\alpha))}| \geq \frac{\hat{\epsilon}}{L}$. Therefore, the distance of two initial points should be greater than $\frac{\hat{\epsilon}}{L}$, otherwise within one Newton step the sequence could jump to a point closer to other initial points and fall into the convergence regions belong to other initial points.

Theorem 8. *When the distribution of X has positive mass on both sides of a, there exists an L such that $|\frac{d}{d\alpha} g_2(\alpha, h(\alpha))| \leq L$ for all α.*

This shows we can find an approximate root of g_2 using a convergent Newton's method and thus have a convergence method to solve the two-moment problem described in this section to approximate optimality given an arbitrary tolerance level.

5.1.2. IFR and IGFR distributions We only discuss how to solve equation $g_2(r_1, h(r_1)) = 0$ for an IGFR distribution, because solving such an equation for an IFR distribution follows a similar logic. Recall that $\alpha_1 = 1$ and $\alpha_2 = \alpha_1 \prod_{w=1}^{v_1-k}(1 - \frac{r_1}{w+k-1})$ for the IGFR distribution. Then, the first and second moment condition become

$$\begin{aligned} g_1(r_1, r_2) :=& (k-q_1)^2\epsilon^2 + \sum_{j=k+1}^{v_1} \prod_{w=1}^{j-k} \left(1 - \frac{r_1}{w+k-1}\right)(j-q_1)^2\epsilon^2 \\ &+ \sum_{j=v_1+1}^{l} \prod_{w=1}^{v_1-k} \left(1 - \frac{r_1}{w+k-1}\right) \prod_{w=1}^{\ell-v_1} \left(1 - \frac{r_2}{w+v_1-1}\right)(j-q_1)^2\epsilon^2 = 1, \end{aligned} \tag{28}$$

$$g_2(r_1,r_2) := (k-q_1)\epsilon + \sum_{j=k+1}^{v_1}\prod_{w=1}^{j-k}\left(1-\frac{r_1}{w+k-1}\right)(j-q_1)\epsilon$$
$$+\sum_{j=v_1+1}^{l}\prod_{w=1}^{v_1-k}\left(1-\frac{r_1}{w+k-1}\right)\prod_{w=1}^{\ell-v_1}\left(1-\frac{r_2}{w+v_1-1}\right)(j-q_1)\epsilon = 0, \tag{29}$$

where $0<r_1,r_2<1$. Obviously, $g_1(r_1,r_2)$ is monotonically decreasing with respect to r_2. Thus, for any given r_1, there is at most a unique choice of r_2 such that $g_1(r_1,r_2)=1$, which further implies a function $h(\cdot)$ with $r_2 = h(r_1)$. As a result, for fixed r_1, we can compute the value of r_2 through the bisection search on the equation $g_1(r_1,r_2)=1$, and the value of $g_2(r_1,h(r_1))$ efficiently. Moreover, the derivative $\frac{\mathrm{d}g_2(r_1,h(r_1))}{\mathrm{d}r_1}$ can also be evaluated by similar approach described earlier for the LC distribution. Therefore, Newton's method is applicable to solve the equation $g_2(r_1,h(r_1))=0$ on the interval [0,1].

5.2. Binomial moment-constrained problem

To illustrate the performance of the proposed computational approach, we implement it on a concrete example that appears in the literature (Subasi et al. 2009). The main focus of the paper is the theoretical properties for global optima of shape-constrained discrete moment problems instead of developing fast algorithms. Therefore, we provide this example only for illustrative purposes. A more in-depth investigation of efficient computation methods for general problems is left for future work.

In Subasi et al. (2009), the authors aim to solve a specific discrete moment problem (Example 4) with the LC constraints in (2). However, their methodology requires relaxing the constraint to be unimodal, which they solve via linear programming. As a type of benchmark, we compare their bounds with the bounds derived from our method.

Our benchmark calculations use the unimodal relaxation of Subasi et al. (2009), described below in our notation. The sample space (before scaling) is always the natural numbers up to $n-1$ (that is, $w_j = j-1$) and we solve

$$\begin{aligned}\max_{K}\ \max_{x\in\mathbb{R}^n}\ & \sum_{j=1}^{n} f_j x_j \\ \text{s.t. } & \sum_{j=1}^{n} w_j^i x_j = q_i \text{ for } i\in[0,2]\end{aligned}$$

$$x_j \le x_{j+1} \text{ for } j \in [1, K-1]$$
$$x_j \ge x_{j+1} \text{ for } j \in [K, n-1],$$

where K is the "mode" of the distribution. Instead of moment constraints, we use the binomial moment constraints of Subasi et al. (2009) with data specified in Table 1; that is,

$$\sum_{j=1}^{n} \binom{w_j}{i} x_j = S_i, \text{ for } i \in [0, 2],$$

where the data S_0, S_1, S_2 can be transformed to moment data q_0, q_1, q_2 via the linear transformation: $q_0 = S_0$, $q_1 = S_1$, and $q_2 = 2S_1 + S_2$. This linear transformation can be extended to higher moments (see (Prékopa 2013, Section 5.6) for details). The objective function is the probability mass $\Pr(X \ge 1) = \sum_{j=1}^{n} f_j x_j = \sum_{j=2}^{n} x_j$ on the positive values of w_j. This objective provides an upper bound on the tail probability given the first two moments. Optimizing the negative of this objective also allows us to calculate lower bounds on tail probabilities. The results are shown in Table 1. We consider three different supports: $[0, 4]$, $[0, 10]$, and $[0, 20]$. The first two were considered in Subasi et al. (2009). The latter is included to demonstrate how our approach scales with the support size.

We also include in the table comparisons to naive methods that solve (DMP-LC) in these instances using the stock global optimization solver in Matlab's optimization toolkit.1. As the reader can see, where the Matlab solver converges, it agrees with the results found using our methodology. Moreover, in many cases, the Matlab solver cannot find an optimal solution.

The LC constraint gives tighter lower and upper bounds in all cases. This finding is expected because the unimodal relaxation is indeed a relaxation. By solving the original LC version of the problem we are able to achieve tighter lower and upper bounds.

5.3. Robust newsvendor problem

We consider the standard newsvendor problem with random demand X distributed by measure μ, overage cost c_o and shortage cost c_s. Order quantity Q is chosen to minimize the expected sum of the two costs:

$$\min_{Q} f(Q; \mu) := c_o \mathbb{E}_\mu [Q - X]^+ + c_s \mathbb{E}_\mu [X - Q]^+.$$

Table 1 Numerical results for the bounds for the total probability for non-negative values with different constraints and methods. The bounds with * in the Naive Method have exit-flags **Local optimal/Infeasible** returned from the toolkit.

			Unimodal			Log-concave Algorithm 1			Log-concave Naïve		
n	S_1	S_2	LB	UB	Time (sec)	LB	UB	Time (sec)	LB	UB	Time (sec)
5	1.9	1.3	0.8750	1	3.2	0.9000	1	3.9	0.9000	1	1.3
5	2.1	1.3	0.9750	1	3.3	0.9920	1	3.4	0.9920	1	0.6
5	1.9	1.7	0.8000	1	3.0	0.8094	0.8433	3.2	0.8096*	0.8232*	0.9
11	5.2	13.1	0.9482	1	7.6	0.9684	1	62.5	0.8699*	1	2.4
11	4.6	13.1	0.8745	1	8.5	0.8924	0.9026	55.2	0.6995*	0.9996*	4.9
11	5.2	15.1	0.9208	1	7.2	0.9310	0.9921	52.9	0.9074*	1*	3.8
21	6.2	28.4	0.79	1	20.0	0.8934	0.97	656.2	0.6214*	1*	5.2
21	15.0	114.5	0.987	1	20.1	0.9956	1	442.5	0.9301*	1*	4.2
21	16.6	132.6	0.9962	1	16.5	0.9997	1	431.6	0.9813*	1*	6.3

Algorithm 2 Solving robust newsvendor problem with LC, IFR, or IGFR shape uncertainty

Input: A lower bound Q_l, an upper bound Q_u and error tolerance δ;
Let $\phi = (1+\sqrt{5})/2$, $Q_1 = Q_u - (Q_u - Q_l)/\phi$ and $Q_2 = Q_l + (Q_u - Q_l)/\phi$;
Compute $f_1 = \tilde{f}(Q_1)$ and $f_2 = \tilde{f}(Q_2)$ by Algorithm 1;
while $Q_2 - Q_1 \geq \delta$ **do**
  if $f_1 \geq f_2$ **then**
    Update $Q_u = Q_2$, $Q_2 = Q_1$ and $f_2 = f_1$;
    Compute $Q_1 = Q_u - (Q_u - Q_l)/\phi$ and $f_1 = f(Q_1)$ by Algorithm 1;
  else
    Update $Q_l = Q_1$, $Q_1 = Q_2$ and $f_1 = f_2$;
    Compute $Q_2 = Q_l + (Q_u - Q_l)/\phi$ and $f_2 = f(Q_2)$ by Algorithm 1;
  end if
end while
Return $(Q_1 + Q_2)/2$.

Given a family of distributions $\mathcal{P}$, the *robust* newsvendor problem is

$$\min_{Q} \max_{\mu \in \mathcal{P}} f(Q; \mu).$$

We further restrict the distributions in $\mathcal{P}$ to have common first- and second-order moments as well as a given shape constraint among LC, IFR, or IGFR:

$$\mathcal{P} = \left\{\mu \text{ is LC, IFR, or IGFR} \mid \mathbb{E}_\mu[X] = q_1,\ \mathbb{E}_\mu[X^2] = q_2\right\}.$$

Note $f(Q;\mu)$ is convex in Q for any μ. Therefore, $\tilde{f}(Q) := \max_{\mu\in\mathcal{P}} f(Q;\mu)$ is convex in Q as well. Moreover, for given Q, computing the value $\tilde{f}(Q)$ amounts to finding the worst-case distribution μ from $\mathcal{P}$, which can be solved by Algorithm 1. Hence, we provide a golden-section search procedure in Algorithm 2 to solve $\min_Q \tilde{f}(Q)$ efficiently.

To test the capability of our approach, we vary the ratios between c_o and c_s to generate different newsvendor problems while fixing the moment values $q_1 = 2.155$ $q_2 = 1.527$ with support $[0, 20]$ to construct the distribution set $\mathcal{P}$. We solve $\min_Q \max_{\mu\in\mathcal{P}} f(Q;\mu)$ with no shape constraint, unimodal constraint, IGFR constraint, IFR constraint, and LC constraint. The last three types of constraints are solved by Algorithm 2. When there is no shape constraint, we have the closed-form solution of Ninh et al. (2018). The problem with the unimodal constraint is modeled as a linear program and solved by the competitive commercial solver *Gurobi*.

The left subfigure in Figure 3 shows the optimal cost curves associated with no shape constraint (the Scarf solution), IGFR constraint, IFR constraint, and LC constraint plotted in a decreasing order, which matches the inclusion relations among the distribution set $\mathcal{P}$ defined by those shape constraints. Therefore, the LC constraint gives the least conservative robust solution, because the set of LC distributions is the most restrictive. The unimodal and the IFR cost curves because an IFR distribution need not be unimodal, as shown in the following example.

Example 1. Consider a distribution μ with support $[1, 4]$. Let $r_1 = 4/5$, $r_2 = 2/3$, and the tail probability $y_1 = 1$, $y_2 = r_1$, $y_3 = r_1 r_2$, and $y_4 = r_1 r_2^2$. Obviously, $\{y_1, y_2\}$ and $\{y_3, y_4\}$ satisfy (13) and constitute the two pieces. Moreover, $y_2^2 = r_1^2 > r_1 r_2 = y_1 y_3$ satisfies (5c) at the break point. Therefore, μ indeed is an IFR distribution. On the other hand, according to the values of the y_i, we have that $x_1 = 1/5$, $x_2 = 12/45$, $x_3 = 8/45$, and $x_4 = 16/45$. Therefore, μ is not unimodal.

Finally, the right subfigure of Figure 3 shows the run time for solving the problem with LC, IFR, and IGFR constraints are consistent across different instances.

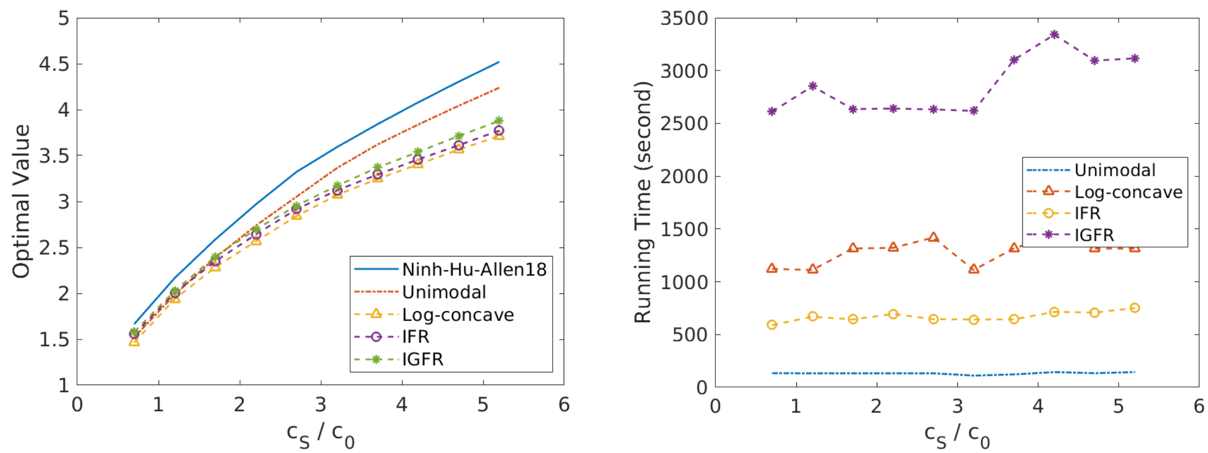

Figure 3 Performance comparison and run time for robust newsvendor model. Recall that c_o is the overage cost and c_s is the shortage cost, so the horizontal axis in both figures in the relative cost c_s/c_o.

6. Conclusion

In summary, we use a reverse convex optimization approach to characterize optimal extreme point distributions for moment problems with reverse convex shape constraints. This characterization allowed us to design an exact low-dimensional algorithm for solving these problems to optimality.

The results in this paper can be applied and built on in several directions, which we leave for future work. First, exploration of additional applications in robust optimization involving shape constraints is one obvious example. Our work on the robust newsvendor problem is a proof of concept of this direction of research. Second, although these results are for the discrete moment problem, we believe they can be extended through limiting arguments to the continuous case. Lastly, future research could more deeply explore implementations of our computational approach paying attention to issues of numerical stability and scaling properties.

Acknowledgments

Xi Chen is supported by the National Science Foundation Faculty Early Career Development Grant (Grant IIS-1845444). Simai He is supported in by National Natural Science Foundation of China under grants 71771141 and 71825003 and the Program for Innovative Research Team of Shanghai University of Finance and Economics. Bo Jiang is supported by the National Natural Science Foundation of China under Grants 11771269 and 11831002 and in part by the Program for Innovative Research Team of Shanghai University of Finance and Economics.

References

Aliprantis C, Border K (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide* (Springer), third edition. 15, ec10

An M (1997) Log-concave probability distributions: Theory and statistical testing. *Duke University Dept of Economics Working Paper* . 10

Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. *Economic Theory* 26(2):445–469. 4

Bagnoli M, Salant SW, Swierzbinski JE (1989) Durable-goods monopoly with discrete demand. *Journal of Political Economy* 97(6):1459–1478. 3

Balabdaoui F, Jankowski H, Rufibach K, Pavlides M (2011) Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society* 75(4):769–790. 22

Banciu M, Mirchandani P (2013) New results concerning probability distributions with increasing generalized failure rates. *Operations Research* 61(4):925–931. 2, 5, 10

Bandi C, Bertsimas D (2012) Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming* 1–48. 2

Barlow R, Proschan F (1996) *Mathematical Theory of Reliability*, volume 17 (SIAM). 4, 12

Baron DP, Myerson RB (1982) Regulating a monopolist with unknown costs. *Econometria* 50(4). 4

Ben-Tal A, Nemirovski A (2001) *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (SIAM). ec2

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Mathematical Programming* 167(2):235–292. 2

Bertsimas D, Natarajan K, Teo CP (2006) Persistence in discrete optimization under data uncertainty. *Mathematical Programming* 108(2):251–274. 3

Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization* 15(3):780–804. 2

Bertsimas D, Tsitsiklis J (1997) *Introduction to Linear Optimization* (Athena). ec5, ec8

Burer S (2009) On the copositive representation of binary and continuous nonconvex quadratic programs. *Mathematical Programming* 120(2):479–495. 6

Canonne CL, Diakonikolas I, Gouleakis T, Rubinfeld R (2018) Testing shape restrictions of discrete distributions. *Theory of Computing Systems* 62(1):4–62. 9, 10, 22

Chakraborty S (2015) Generating discrete analogues of continuous probability distributions-a survey of methods and constructions. *Journal of Statistical Distributions and Applications* 2(1):6. 5

Chen Z, Sim M, Xu H (2018) Distributionally robust optimization with infinitely constrained ambiguity sets, Working Paper. 2

Dai T, Jerath K (2016) Impact of inventory on quota-bonus contracts with rent sharing. *Operations Research* 64(1):94–98. 4

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612. 2

Dennis Jr J, Schnabel R (1996) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 (SIAM). 29

Duembgen L, Huesler A, Rufibach K (2011) Active set and EM algorithms for log-concave densities based on complete and censored data. *arXiv preprint arXiv:0707.4643v4* . 4

Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050* . 2

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199* . 2

Gavirneni S, Kapuscinski R, Tayur S (1999) Value of information in capacitated supply chains. *Management Science* 45(1):16–24. 4

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Operations Research* 58(4-part-1):902–917. 2

Han Q, Du D, Zuluaga LF (2014) Technical note: A risk and ambiguity-averse extension of the max-min newsvendor order formula. *Operations Research* 62(3):535–542. 3

Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *Operations Research* 66(3):849–869. 6

Hanasusanto GA, Kuhn D, Wallace SW, Zymler S (2015) Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming* 152(1):1–32. 3

He S, Zhang J, Zhang S (2010) Bounding probability of small deviation: A fourth moment approach. *Mathematics Operations Research* 35(1):208–232. 2

Hillestad R, Jacobsen S (1980) Reverse convex programming. *Applied Mathematics & Optimization* 6(1):63–78. 6, 16, 17

Horst R, Thoai N (1999) DC programming: Overview. *Journal of Optimization Theory and Applications* 103(1):1–43. 6

Jiang R, Wang J, Guan Y (2012) Robust unit commitment with wind power and pumped storage hydro. *IEEE Transactions on Power Systems* 27(2):800–810. 2

Johnson O, Goldschmidt C (2006) Preservation of log-concavity on summation. *ESAIM: Probability and Statistics* 10:206–215. 22

Johnson O, Kontoyiannis I, Madiman M (2013) Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound poisson measures. *Discrete Applied Mathematics* 161(9):1232–1250. 22

Klee V (1957) Extremal structure of convex sets. *Archiv der Mathematik* 8(3):234–240. ec4

Laffont JJ, Tirole J (1988) The dynamics of incentive contracts. *Econometria* 56(5):1153–1175. 4

Lam H, Mottet C (2017) Tail analysis without parametric models: A worst-case perspective. *Operations Research* 65(6):1696–1711. 3

Lariviere M (2006) A note on probability distributions with increasing generalized failure rates. *Operations Research* 54(3):602–604. 5

Lewis TR, Sappington DEM (1988) Regulating a monopolist with unknown demand. *American Economic Review* 78(5):986–998. 4

Li B, Jiang R, Mathieu JL (2017) Ambiguous risk constraints with moment and unimodality information. 3

Long DZ, Qi J (2014) Distributionally robust discrete optimization with entropic value-at-risk. *Operations Research Letters* 42(8):532–538. 2

Matthews S (1987) Comparing auctions for risk averse buyers: A buyer's point of view. *Econometrica* 55(3):633–646. 4

Meyer R (1970) The validity of a family of optimization methods. *SIAM Journal on Control* 8(1):41–54. 6

Myerson RB, Satterthwaite MA (1983) Efficient mechanisms for bilateral trading. *Journal of Economic Theory* 29:265–281. 4

Natarajan K, Sim M, Uichanco J (2017) Asymmetry and ambiguity in newsvendor models. *Management Science* 64(7):3146–3167. 2, 3

Natarajan K, Teo CP, Zheng Z (2011) Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Operations research* 59(3):713–728. 6

Ninh A, Hu H, Allen D (2018) Robust newsvendor problems: effect of discrete demands. *Annals of Operations Research* 1–15. 34

Ninh A, Hu H, Allen D (2019) Robust newsvendor problems: effect of discrete demands. *Annals of Operations Research* 275(2):607–621. 7

Peña J, Vera J, Zuluaga L (2015) Completely positive reformulations for polynomial optimization. *Mathematical Programming* 151(2):405–431. 6

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Operations Research* 56(1):188–203. 3

Popescu I (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research* 30(3):632–657. 2, 4, 8

Prékopa A (1990) Sharp bounds on probabilities using linear programming. *Operations Research* 38(2):227–239. 3, 9

Prékopa A (2013) *Stochastic Programming*, volume 324 (Springer Science & Business Media). 2, 32

Prékopa A, Boros E (1991) On the existence of a feasible flow in a stochastic transportation network. *Operations Research* 39(1):119–129. 3

Prékopa A, Long J, Szantai T (2004) New bounds and approximations for the probability distribution of the length of the critical path. *Dynamic Stochastic Optimization*, 293–320 (Springer). 3

Prékopa A, Ninh A, Alexe G (2016) On the relationship between the discrete and continuous bounding moment problems and their numerical solutions. *Annals of Operations Research* 238(1-2):521–575. 3

Rujeerapaiboon N, Kuhn D, Wiesemann W (2018) Chebyshev inequalities for products of random variables. *Mathematics of Operations Research* 43(3):887–918. 2

Saghafian S, Tomlin B (2016) The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research* 64(1):167–185. 2, 3

Saumard A, Wellner JA (2014) Log-concavity and strong log-concavity: a review. *Statistics Surveys* 8:45–114. 22

Scarf H, Arrow K, Karlin S (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production* 10:201–209. 2

Subasi E, Subasi M, Prékopa A (2009) Discrete moment problems with distributions known to be unimodal. *Mathematical Inequalities and Applications* 12(3):587–610. 4, 5, 31, 32

Sudheesh K, Anisha P, Deemat C (2015) A simple approach for testing constant failure rate against different ageing classes for discrete data. *Statistical Methodology* 25:74–88. 5

Swaminathan JM, Shanthikumar JG (1999) Supplier diversification: effect of discrete demand. *Operations research letters* 24(5):213–221. 2, 3

Tian R, Cox SH, Zuluaga LF (2017) Moment problem and its applications to risk assessment. *North American Actuarial Journal* 21(2):242–266. 2

Walther G (2000) Inference and modeling with log-concave distributions. *Statistical Science* 24(3). 4

Wellner JA (2013) Strong log-concavity is preserved by convolution. *High Dimensional Probability VI*, 95–102 (Springer). 22

Xu G, Burer S (2018) A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides. *Computational Optimization and Applications* 70(1):33–59. 6

Ziya S, Ayhan H, Foley RD (2004) Relationships among three assumptions in revenue management. *Operations Research* 52(5):804–809. 5

Proofs

EC.1. Proof of Proposition 1

Setting $u=v=1$ in (3c) specializes to (2c). Further, (3c) guarantees consecutive support: if there exist $j_1<j_2<j_3$ such that $x_{j_1},x_{j_3}>0$, $x_{j_2}=0$, by setting $u=j_2-j_1, v=j_3-j_2$, the constraint $x_{j-u}^{v}x_{j+v}^{u}\le x_j^{u+v}$ for $j=j_2$ is violated. Hence, every feasible distribution of (DMP-LC') is a feasible distribution of (DMP-LC) with the same objective value (note the objectives of both problems are identical).

On the other hand, any feasible distribution to problem (2) with support $[k,\ell]$ satisfies (3c), and by a straightforward induction starting with (2c) as a base case. To be specific, we assume that for some u and v, $x_{j-u}^{v}x_{j+v}^{u}\le x_j^{u+v}$ holds for $j\in(k,\ell), j-u\ge k, j+v\le \ell$, and we want to show that $x_{j-u}^{v+1}x_{j+v+1}^{u}\le x_j^{u+v+1}$. Then it suffices to establish

$$x_{j-u}x_{j+v+1}^{u}\le x_{j+v}^{u}x_j, \tag{EC.1}$$

From (2c), we have that $x_{j+v}^2\ge x_{j+v-1}x_{j+v+1}$, $x_{j+v-1}^2\ge x_{j+v-2}x_{j+v},\cdots,x_{j-k}^2\ge x_{j-k+1}x_{j-k-1}$. Putting those equalities together by applying multiplications to and canceling out the same terms on both sides give that

$$x_{j+v}x_{j-k}\ge x_{j+v+1}x_{j-k-1}, \quad 0\le k\le j-2. \tag{EC.2}$$

Hence, starting from $x_{j+v}x_j\ge x_{j+v+1}x_{j-1}$, multiplying it by (EC.2) with $1\le k\le u-1$, and (EC.1) follows as required by canceling out the same terms on both sides.

For those points j such that $j-u$ or $j+v$ is outside the support, or the middle point j outside the support, the constraint $x_{j-u}^{v}x_{j+v}^{u}\le x_j^{u+v}$ holds naturally because the left-hand side is zero for these cases. In other words, (3c) is satisfied. Hence, every feasible distribution of (DMP-LC) is a feasible distribution of (DMP-LC') with the same objective value. $\square$

EC.2. Proof of Proposition 2

We first prove a preliminary lemma for establishing Proposition 2.

LEMMA EC.1. *The set* $\{(x,y,z): x^u y^v \geq z^{u+v}, x\geq 0, y\geq 0, z\geq 0\}$ *is convex for any positive integers* u *and* v. *Moreover, the set* $\{(x,y,z): x^u y^v \geq (z+\epsilon)^{u+v}, x\geq 0, y\geq 0, z\geq 0\}$ *is also convex for any positive integers* u *and* v *and nonnegative real* ϵ.

Proof of Lemma EC.1. For any integers u and v, let t be an integer such that $u+v\leq 2^t$. From point 11 on page 95 of Ben-Tal and Nemirovski (2001)), the set $\{(x_1,\cdots,x_{2^t},z): x_j\geq 0,\ 1\leq j\leq 2^t,\ z\leq (\Pi_{j=1}^{2^t} x_j)^{1/2^t}\}$ is conic-quadratic representable, and thus convex. Therefore, when intersecting with linear constraints, the set

$$\begin{aligned}\{(x_1,\ldots,x_{2^t},z): x_j\geq 0,\ 1\leq j\leq 2^t,\ 0\leq z\leq (\Pi_{j=1}^{2^t} x_j)^{1/2^t},\ x_j=x_k \text{ for } j,k\in[1,u]\\ x_j=x_k \text{ for } j,k\in[u+1,u+v],\ \text{and } x_j=z \text{ for } j,k\in[u+v+1,2^t]\}\end{aligned} \tag{EC.3}$$

is also convex. Now, consider the image of this set under the coordinate projection onto x_1 (relabeled as x), x_{u+1} (relabeled as y), and z. The set in (EC.3) becomes

$$S := \{(x,y,z): z\leq (x^u y^v z^{(2^t-u-v)})^{1/2t}, x\geq 0, y\geq 0, z\geq 0\},$$

which is convex because the projection of a convex set is convex. Observe that the constraint $z\leq (x^u y^v z^{(2^t-u-v)})^{1/2t}$ is equivalent to $z^{u+v}\leq x^u y^v$ and so the set S is precisely the set in the statement of the lemma. The fact S is a convex set proves the result.

To show the "moreover", simply define $w=z+\epsilon$ and the set in question can be written $T=\{(x,y,w): x^u y^v\geq w^{u+v}, x\geq 0, y\geq 0, w\geq \epsilon\}$. Note that this set is the intersection of $\{(x,y,w): x^u y^v\geq w^{u+v}, x\geq 0, y\geq 0, w\geq 0\}$ (convex by the previous part of the lemma) and the convex set $\{(x,y,w): w\geq\epsilon\}$. Thus, T is convex because the intersection of convex sets is convex. $\square$

Proof of Proposition 2. First, define the set

$$S(\epsilon)\triangleq \{(x,y,z): x^u y^v\geq (z+\epsilon)^{u+v},\ x\geq 0, y\geq 0, z\geq 0\}.$$

By Lemma EC.1, $S(\epsilon)$ is convex for every $\epsilon>0$ and the sets are nested; that is, $S(\epsilon_1)\subseteq S(\epsilon_2)$ for $\epsilon_1\geq\epsilon_2$, which implies the union $\bigcup_{\epsilon>0} S(\epsilon)$ is also a convex set (the union of convex nested sets is convex).

Next, we make the following claim:

$$S \triangleq \{(x,y,z) : x^u y^v > z^{u+v},\ x \geq 0, y \geq 0, z \geq 0\} = \bigcup_{\epsilon>0} S(\epsilon). \tag{EC.4}$$

This claim suffices to prove the result because $\bigcup_{\epsilon>0} S(\epsilon)$ is convex.

The $\supseteq$ containment in (EC.4) holds because $S(\epsilon) \subseteq S$ for all $\epsilon > 0$. Conversely, to show $\subseteq$ in (EC.4), let $(x,y,z) \in S$. Hence, we must have $x^u y^v - z^{u+v} = \delta > 0$. Next, we show there exists an $\epsilon > 0$ such that $x^u y^v = (z+\epsilon)^{u+v}$, in which case, $(x,y,z) \in S(\epsilon)$ and we are done. Taking $\epsilon = (z^{u+v} + \delta)^{\frac{1}{u+v}} - z$ does the trick. $\square$

EC.3. Proof of Lemma 1

According to Definition 2, we have the following inequality if x is an IFR distribution:

$$\frac{x_j}{\sum_{k=j}^n x_k} - \frac{x_{j+1}}{\sum_{k=j+1}^n x_k} \leq 0, \text{ for } j \in [1, n-1],$$

which is equivalent to

$$x_j \sum_{k=j+1}^n x_k - x_{j+1} \sum_{k=j}^n x_k \leq 0, \text{ for } j \in [1, n-1], \tag{EC.5}$$

whereas if $\{\bar{F}_1, \ldots, \bar{F}_n\}$ is log-concave, we have

$$\sum_{k=j-1}^n x_k \sum_{k=j+1}^n x_k - \left(\sum_{k=j}^n x_k\right)^2 \leq 0, \text{ for } j \in [2, n-1],$$

which is equivalent to

$$x_j \sum_{k=j+1}^n x_k - x_{j+1} \sum_{k=j}^n x_k \leq 0, \text{ for } j \in [1, n-2]. \tag{EC.6}$$

Inequalities (EC.5) and (EC.6) are the same except that (EC.6) does not include the case in which $j = n-1$. In this case, the inequality holds naturally: $x_{n-1}x_n - x_n(x_{n-1} + x_n) \leq 0$. Thus, the two definitions of IFR distribution are equivalent. $\square$

EC.4. Proof of Proposition 3

First, define the set

$$S(\epsilon) \triangleq \left\{ (x,y,z) : \left\| \begin{pmatrix} ax \\ by+cz \end{pmatrix} \right\| \leq ax + cz - \epsilon \right\}.$$

Obviously, $S(\epsilon)$ is a standard SOC and thus convex for every $\epsilon > 0$. Moreover, the sets are nested as $S(\epsilon_1) \subseteq S(\epsilon_2)$ for $\epsilon_1 \geq \epsilon_2$, which implies that the union $\bigcup_{\epsilon>0} S(\epsilon)$ is also a convex set, because the union of convex nested sets is convex. Next, we make the following claim:

$$S \triangleq \left\{ (x,y,z) : \left\| \begin{pmatrix} ax \\ by+cz \end{pmatrix} \right\| < ax + cz \right\} = \bigcup_{\epsilon>0} S(\epsilon). \tag{EC.7}$$

This claim suffices to prove the result because $\bigcup_{\epsilon>0} S(\epsilon)$ is convex. The $\supseteq$ containment in (EC.7) holds since $S(\epsilon) \subseteq S$ for all $\epsilon > 0$. Conversely, to show $\subseteq$ in (EC.7), let $(x,y,z) \in S$. Hence, we must have

$$ax + cz - \left\| \begin{pmatrix} ax \\ by+cz \end{pmatrix} \right\| = \delta > 0.$$

Taking $\epsilon = \delta$, we have $(x,y,z) \in S(\epsilon)$. □

EC.5. Proof of Theorem 1 and Lemma 4

Proof of Lemma 4. The fact that $\operatorname{ext} \overline{\operatorname{conv}}\, S \subseteq S$ follows immediately from Klee (1957, Theorem 3.5). Suppose, by way of contradiction, that there exists an $x \in \operatorname{ext} \overline{\operatorname{conv}}\, S$ that is not an extreme point of S. Then, there exist $y, z \in S$ with $y \neq z$ such that $x = \lambda y + (1-\lambda) z$ where $\lambda > 0$. However, since $y, z \in \overline{\operatorname{conv}}\, S$, this contradicts that $x \in \operatorname{ext} \overline{\operatorname{conv}}\, S$. The result then holds. □

With Lemma 2, Lemma 3, and Lemma 4 in hand, we can now establish Theorem 1.

Proof of Theorem 1. The problem $\min\{c(x) : x \in \overline{\operatorname{conv}}\, S\}$ has an optimal extreme point solution $x^* \in \operatorname{ext} \overline{\operatorname{conv}}\, S$ by Lemma 2 and the fact that $\overline{\operatorname{conv}}\, S$ is a compact convex set by Lemma 3. Since $S \subseteq \overline{\operatorname{conv}}\, S$, $\min\{c(x) : x \in \overline{\operatorname{conv}}\, S\} \leq \min\{c(x) : x \in S\}$. However, since $x^* \in S$, by Lemma 4, we have $c(x^*) = \min\{c(x) : x \in \overline{\operatorname{conv}}\, S\} \leq \min\{c(x) : x \in S\} \leq c(x^*)$ since x^* is optimal to the minimization over $\overline{\operatorname{conv}}\, S$ and feasible to the minimization over S. However, all inequalities must therefore be equalities and so $\min\{c(x) : x \in S\} = c(x^*)$. Since $x^* \in \operatorname{ext} S$ by Lemma 4, (8) must have an optimal extreme-point solution. □

EC.6. Proof of Lemma 5

Since c is lower-semicontinuous and quasiconcave and F is compact, Theorem 1 implies that there exists an optimal extreme point solution x^*. Let F^* denote the connected component of F that contains x^* and restrict attention to those R_p that intersect F^*. Let $C_p = \mathbb{R}^n \setminus R_p$. Then, C_p is an open convex set since R_p is closed and reverse convex. Since $x^* \in F$, $x^* \notin C_p$ for all p. For all p, let y_p be such that $\text{dist}(x^*, \text{cl}(C_p)) = \text{dist}(x^*, y_p)$; that is, y_p minimizes the distance between x^* and the closure of C_p.

Since x^* is an extreme point of F, there exists a p such that $x^* \in \text{bd}(R_p)$. By definition, $\text{bd}(C_p) = \text{bd}(R_p)$ and so there exists a p such that $x^* \in \text{bd}(C_p)$. Clearly, if $x^* \in \text{bd}(C_p)$ then $\text{dist}(x^*, \text{cl}(C_p)) = 0$ and so $y_p = x^*$. Hence, there always exists a p such that $y_p = x^* \in \text{bd}(C_p)$.

Using the vector y_p, we can define for all $p \in [1, P]$ a supporting hyperplane of $\text{cl}(C_p)$ with normal α_p and right-hand side β_p that weakly separates C_p from the point x^*. These hyperplanes define the polyhedron $\hat{F} = \{x : \alpha_p^\top x \geq \beta_p, \text{ for } p \in [1, P]\}$ so that $\alpha_p^\top x^* \geq \beta_p$ for all p. Such a choice is always possible since each hyperplane weakly separates C_p from the point x^*, which implies $x^* \in \hat{F}$. In the special case in which $y_p \neq x^*$, the hyperplane $\{x : (x^* - y_p)^\top (x - y_p) \leq 0\}$ does the trick, by the standard projection theorem. Next, we claim $\hat{F} \subseteq F^*$. Indeed, since $\alpha_p^\top x \leq \beta_p$ is a supporting hyperplane of $\text{cl}(C_p)$, the set of x that satisfies $\alpha_p^\top x \geq \beta_p$ lies on the boundary of C_p or outside of C_p. Such an x lies entirely inside of R_p, which implies $\hat{F}$ is a subset of R_p for all p, and so $\hat{F} \subseteq F^*$.

Consider the optimization problem

$$\begin{aligned} &\min\ c(x) \\ &\ s.t.\ \alpha_p^\top x \geq \beta_p \text{ for } p \in [1, P]. \end{aligned} \tag{EC.8}$$

Since $x^* \in \hat{F} \subseteq F^*$ and x^* is an optimal solution of the original problem, x^* is an optimal solution of (EC.8). Moreover, x^* is an extreme point of $\hat{F}$ and so at least n linearly independent constraints at x^* are tight, by the characterization of extreme points of polyhedra (Bertsimas and Tsitsiklis 1997, Theorem 2.3). Hence, at least n of the inequalities $\alpha_p^\top x \geq \beta_p$ must be tight at $x = x^*$. The points in $\hat{F}$ that satisfy $\alpha_p^\top x = \beta_p$ are boundary points of C_p since $\alpha_p^\top x = \beta_p$ is a supporting hyperplane of the convex set C_p. Since $\text{bd}(R_p) = \text{bd}(C_p)$, the points in $\hat{F}$ that satisfy $\alpha_p^\top x = \beta_p$ are boundary points of R_p. Hence, x^* lies on the boundary of at least n of the sets R_p. $\square$

EC.7. Proof of Theorem 2

This proof uses the following notation. For any feasible solution x to (Rev-Cvx), let $S(x)$ denote the support of x; that is, $S(x) = \{j : x_j > 0\}$. Let A^i denote the i-th row of the matrix A and A_j the j-th column. For any subset S of $[1,n]$ (e.g., the support of a feasible solution), let $A_S = [A_j]_{j\in S}$. That is, A_S is the submatrix of A consisting of the columns indexed by S. Note A_S^r denotes the r-th row of the matrix A_S and let $\mathcal{L}(S) = \text{span}(\{(A_S^1)^\top, \ldots, (A_S^m)^\top\}$ denote the span of the rows of A_S. Finally, let $\nabla f_p(x)$ denote the gradient of f_p at x, where $[\nabla f_p(x)]_S$ is the gradient of f_p restricted to the components in the subset S.

Now, for the proof. Since c is continuous and quasiconcave and F is a compact set, by Theorem 1 there exists an optimal extreme point solution. For any such optimal extreme point x^* with support $S = S(x^*)$, define

$$F_0 := \{x : Ax = b,\ x_j = 0 \text{ for } j \notin S,\ x_j > 0 \text{ for } j \in S\}.$$

Then, the feasible region F includes $\{x : f_p(x) \le 0,\ p = 1, \ldots, P\} \cap X_0$. Let $\delta_1 = \min\{x_j^* : x_j^* > 0\}$ and denote

$$F(\delta_1) := \{x : Ax = b,\ x_j = 0 \text{ for } j \notin S,\ x_j \ge \delta_1/2, \text{ for } j \in S\}.$$

Our goal is as follows. For $p = 1, \ldots, P$, we want to construct sets $\hat{F}_p$ of the form

$$\hat{F}_p := \{x : \alpha_p^\top (x - x^*) \le \beta_p\} \cap X(\delta_1), \tag{EC.9}$$

such that

$$\hat{F} := \cap_{p=1}^P \hat{F}_p = \{x : \alpha_p^\top (x - x^*) \le \beta_p,\ p = 1, \ldots, P\} \cap F(\delta_1) \tag{EC.10}$$

is a subset of F, where α_p and $\beta_p \ge 0$ will be specified later. Since $\beta_p \ge 0$, $x^* \in \hat{F}_p$ for all p, and is thus in $\hat{F}$. So, as long as $\hat{F} \subseteq F$ since x^* is an extreme point of F, it must also be an extreme point of $\hat{F}$. Note $\hat{F}$ is defined by linear equalities and inequalities and so there must exist n of them that are tight at point x^*, and we can further check which constraints are tight.

$0 \notin \mathrm{cl}(Y_p)$	$\alpha_p = \hat{\alpha}_p$, $\beta_p = \hat{\beta}_p/2$, $\hat{\alpha}_p$ and $\hat{\beta}_p$ obtained by strong separation
$0 \in \mathrm{cl}(Y_p)$	$\alpha_p = \hat{\alpha}_p$, $\beta_p = 0$, $\hat{\alpha}_p$ are obtained weak separation
$0 \in \mathrm{cl}(Y_p)$, $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$	$\alpha_p = \nabla f_p(x^*)$ and $\beta_p = 0$

Table EC.1 Specifying α_p and β_p in $\hat{X}_p$.

We now construct the $\hat{F}$ in (EC.10). Let $x_S = [x_j]_{j\in S}$ and

$$F_p = F_0 \cap \{(x_S;0) : f_p(x_S;0) > 0\}. \tag{EC.11}$$

Here, $(x_S;0)$ denotes a vector with value 0 outside of the index set S. A key property of F_p is that it admits a strong separation property useful for our arguments (see Claim EC.1 below). To describe this property, we explore a related set in a smaller subspace. Construct matrix $B \in \mathbb{R}^{|S|\times(|S|-\mathrm{rank}(A_S))}$ such that its columns span the whole null space of A_S. That is, $A_S B = 0$ and $\mathrm{rank}(B) = |S| - \mathrm{rank}(A_S)$. Then, we have that

$$\{(x_S;0) : A_S\, x_S = b\} = \{(By + x_S^*;0) : y \in \mathbb{R}^{(|S|-\mathrm{rank}(A_S))}\}. \tag{EC.12}$$

Letting

$$Y_p := \{y : By + x_S^* > 0, f_p(By + x_S^*;0) > 0\},$$

we can define the "strong separation" property of F_p as follows:

Claim EC.1. *(Strong separation) For all p, there exist $\alpha_p^\top$ and $\hat{\beta}_p > 0$ such that*

$$\begin{cases} \hat{\alpha}_p^\top (x - x^*) \ge \hat{\beta}_p > 0, & \text{for } x \in F_p \quad \text{if} \quad 0 \notin \mathrm{cl}(Y_p) \\ \hat{\alpha}_p^\top (x - x^*) > 0, & \text{for } x \in F_p \quad \text{if} \quad 0 \in \mathrm{cl}(Y_p). \end{cases} \tag{EC.13}$$

Moreover, if we further assume $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$ then $\nabla f(x^*)^\top (x - x^*) > 0$ for all $x \in F_p$.

The proof of this claim is somewhat technical and so is relegated to the end of this section. We take it as given and return to constructing $\hat{F}$. According to (EC.10), it suffices to show how to construct $\hat{F}_p$ such that

$$\hat{F}_p \subseteq \{x : f_p(x) \le 0\},\ p \in [1, P], \tag{EC.14}$$

since $F(\delta_1) \subseteq F_0$. In other words, we need to prove that $x \in \hat{F}_p$ implies $f_p(x) \le 0$.

We show (EC.14) in two cases: (i) $0 \notin \mathrm{cl}(Y_p)$ and (ii) $0 \in \mathrm{cl}(Y_p)$. We use Table EC.1 to track some of the notation and details.

In case (i), according to Claim EC.1, there exist a $\hat{\alpha}_p \neq 0$ and $\hat{\beta}_p \neq 0$ such that

$$\hat{\alpha}_p^\top (x - x^*) \geq \hat{\beta}_p > 0 \text{ for all } x \in F_p. \tag{EC.15}$$

By letting $\alpha_p = \hat{\alpha}_p$ and $\beta_p = \frac{\hat{\beta}_p}{2}$, one has $x^* \in \hat{F}_p \neq \emptyset$. Moreover, from definition (EC.9) of $\hat{F}_p$, any $x \in \hat{F}_p$ satisfies $x \in F(\delta_1) \subseteq F_0$ and

$$\hat{\alpha}_p^\top (x - x^*) = \alpha_p^\top (x - x^*) \leq \beta_p = \hat{\beta}_p/2.$$

Combining this statement with (EC.15) yields that $x \notin F_p$ for any $x \in \hat{F}_p$. Then, according to (EC.11), such x does not belong to F_p, simply because it violates the constraint $f_p(x) > 0$. Therefore, we can conclude $f_p(x) \leq 0$ for all $x \in \hat{F}_p$.

In case (ii), again by Claim EC.1, we have $\hat{\alpha}^\top (x - x^*) > 0$ for $x \in F_p$, where $\hat{\alpha} = \nabla f(x^*)$ if $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$. Then, we can take $\alpha_p = \nabla f_p(x^*)$ and $\beta_p = 0$ in (EC.9). Obviously, $x^* \in \hat{F}_p \neq \emptyset$ and $x \notin \hat{F}_p$ for any $x \in F_p$. Similarly, we can argue such an x does not belong to $\hat{F}_p$ due to the violation of the constraint $f_p(x) > 0$. Then, it follows that $f_p(x) \leq 0$ for all $x \in \hat{F}_p$.

So far, we have constructed $\hat{F}_p$ in the form of (EC.11) (as in Table EC.1) and $\hat{F}$ based on (EC.10). Moreover, we have shown $\hat{F} \subseteq F$. Since x^* is an extreme point of F and lies both in F and $\hat{F}$, it is an extreme point of $\hat{F}$ as well. Note $\hat{F}$ is defined by a number of linear equalities and inequalities, then there must exist n of them that are tight and linear independent at point x^* by standard theory (see e.g. Bertsimas and Tsitsiklis (1997, Theorem 2.3)).

Since A is an m by n matrix of rank m, there are $n - m$ tight constraints from

$$\begin{aligned} \alpha_p^\top (x - x^*) \leq \beta_p & \text{ for } p \text{ such that } 0 \notin \mathrm{cl}(Y_p) \\ {\alpha_p}^\top (x - x^*) \leq 0 & \text{ for } p \text{ such that } 0 \in \mathrm{cl}(Y_p) \\ x_j = 0 & \text{ for } j \notin S \\ x_j \geq \delta_1/2 & \text{ for } j \in S, \end{aligned}$$

where $\alpha_p = \nabla f_p(x^*)$ if $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$. Now, we investigate which of the above constraints are tight. First, obviously $x_j^* = 0$ is tight for all $j \notin S$, and $x_j^* \geq \delta_1 > \delta_1/2$ can not be tight for all $j \in S$. Then, for the constraint p such that $0 \notin \mathrm{cl}(Y_p)$ since $\beta_p > 0$, $\alpha_p^\top (x^* - x^*) = 0 < \beta_p$ cannot be tight. Finally, recall we have proved in the previous discussion that $f_p(x^*) \geq 0$ for all p such that $0 \in \mathrm{cl}(Y_p)$. That is, when $f_p(x^*) < 0$, it holds that

$0 \notin \mathrm{cl}(Y_p)$, and thus the corresponding constraint $\alpha_p^\top(x - x^*) \le \beta_p$ cannot be tight at x^*. In summary, all $n - m$ tight constraints come from

$$\begin{aligned} {\alpha_p}^\top(x - x^*) \le 0 \text{ for } p \text{ such that } f_p(x^*) = 0 \text{ and } 0 \in \mathrm{cl}(Y_p) \\ x_j = 0 \text{ for } j \notin S, \end{aligned} \tag{EC.16}$$

which implies $n - m$ of the inequalities

$$\begin{aligned} f_p(x^*) \le 0 \text{ for } p \in [1, P] \\ x_j^* \ge 0 \text{ for } j \in [1, n] \end{aligned}$$

in (Rev-Cvx) are tight. This completes the proof of Theorem 2. □

The proof of Claim EC.1 relies on the following two subclaims.

Subclaim 1 *The set Y_p is a convex and open set for $p = 1, \dots, P$.*

Proof of Subclaim 1: By assumption, $S_1 := \{x : f_p(x) > 0, x \ge 0\}$ is convex. Therefore, $S_2 := \{x : f_p(x) > 0, x \ge 0\} \cap \{x : Ax = b\}$ is also a convex set since we are intersecting S_1 with the convex set $\{x : Ax = b\}$. Moreover, the set $S_3 := S_2 \cap \{x : x_S > 0, x_{\bar{S}} = 0\}$ is again convex since $\{x : x_S > 0, x_{\bar{S}} = 0\}$ is a convex set. Finally, consider the affine map $y \mapsto (By + x_S^*, 0)$. Note that Y_p is the inverse image of this map and therefore convex.

Moreover, for any $y_1 \in Y_p$, let

$$0 < \delta = \min_{j \in \{1, \dots, |S|\}} \{(By_1 + x_S^*)_j : f_p(By_1 + x_S^*; 0) > 0\}.$$

Since $f_p(\cdot)$ is continuous, there exists an $\epsilon > 0$ such that for any y with $\|y - y_1\|_2 \le \epsilon$, we have

$$\min_{j \in \{1, \dots, |S|\}} \{(By + x_S^*)_j : f_p(By + x_S^*; 0) > 0\} \ge \delta/2 > 0.$$

Thus, $y \in Y_p$ and Y_p is open. This completes the proof of Subclaim 1. □

Moreover, we have a “strong separation property” of Y_p described as follows.

Subclaim 2 *There exist a $d_p \ne 0$ and $\hat{\beta}_p > 0$ such that*

$$\begin{cases} d_p^\top y \ge \hat{\beta}_p > 0, \text{ for } y \in Y_p \text{ if } & 0 \notin \mathrm{cl}(Y_p) \\ d_p^\top y > 0, \text{ for } y \in Y_p \text{ if } & 0 \in \mathrm{cl}(Y_p). \end{cases} \tag{EC.17}$$

Moreover, letting $g_p(y) = f_p(By + x_S^*; 0)$ and assuming $\nabla g_p(0) \ne 0$, if $0 \in \mathrm{cl}(Y_p)$ then $\nabla g_p(0)^\top y > 0$ for all $y \in Y_p$.

Proof of Subclaim 2: Note that $f_p(B\cdot 0+x_S^*;0)=f_p(x_S^*;0)=f_p(x^*)\le 0$, thus $0\notin Y_p$. Since, by Subclaim 1, Y_p is convex, $\mathrm{cl}(Y_p)$ is both closed and convex. Then if $0\notin \mathrm{cl}(Y_p)$, by the strong separation theorem for closed convex sets (see, for instance, Aliprantis and Border (2006, Corollary 5.80)), there exist $d_p\neq 0$ and $\hat\beta_p>0$ such that $d_p^\top y\ge \hat\beta_p>0$ for $y\in Y_p$. In the case of $0\in \mathrm{cl}(Y_p)$, weak separation holds. That is, there exists an $\hat\alpha_p\neq 0$ such that

$$\hat\alpha_p^\top y \ge 0^\top y = 0 \text{ for all } y\in Y_p. \tag{EC.18}$$

Since Y_p is open, that weak separation becomes strict; that is, $\hat\alpha_p^\top y>0$ for all y. Otherwise, if there exists a $y'\in Y_p$ such that $\hat\alpha_p^\top y'=0$ then since Y_p is open there exists a $y''\in Y_p$ in a small neighborhood of y' such that $\hat\alpha_p^\top y''<0$. This violates the condition shown above that $\hat\alpha_p^\top y\ge 0$ for $y\in Y_p$. Together this yields (EC.17).

To establish the "moreover" part, note $g_p(y)=f_p(By+x_S^*;0)\ge 0$ for any $y\in \mathrm{cl}(Y_p)$. Hence, $g_p(0)=f_p(x_S^*;0)=f_p(x^*)\le 0$. Combining these two facts gives that $g_p(0)=0$ when $0\in \mathrm{cl}(Y_p)$. That is, 0 is a global minimizer of the problem

$$\begin{array}{l} \min\ g_p(y) \\ \text{s.t.}\ \ y\in \mathrm{cl}(Y_p). \end{array}$$

Thus, the following optimality condition in the form of variational inequality holds: $\nabla g_p(0)^\top(y-0)\ge 0$ for $y\in \mathrm{cl}(Y_p)$, which trivially leads to $\nabla g_p(0)^\top y\ge 0$ for $y\in Y_p$. Since Y_p is open, we get strict separation $\nabla g_p(0)^\top y>0$ for all $y\in Y_p$. This completes the proof of Subclaim 2. $\square$

Proof of Claim EC.1: We show (EC.13) holds with $\hat\alpha_p=(B(B^\top B)^{-1}d_p;\gamma_p)$, with d_p being defined in Subclaim 2 and any $\gamma_p\in\mathbb{R}^{n-|S|}$ and $\hat\beta_p$ as constructed in Subclaim 2. Indeed, for any $x\in F_p$, due to (EC.12), we can find a $y\in Y_p$ such that $x=(By+x_S^*;0)=(By;0)+x^*$. Consequently,

$$\hat\alpha_p^\top(x-x^*)=d_p^\top(B^\top B)^{-\top}B^\top By+\gamma_p^\top 0=d_p^\top y.$$

Then, according to Subclaim 2, (EC.13) holds.

To establish the "moreover" of Claim EC.1, observe that when $\nabla g_p(0)\neq 0$ and $0\in \mathrm{cl}(Y_p)$, by letting $\hat\alpha_p=(B(B^\top B)^{-1}\nabla g_p(0);\gamma_p)$ with any $\gamma_p\in\mathbb{R}^{n-|S|}$, we have $\hat\alpha_p^\top(x-x^*)>0$ for $x\in F_p$. The argument here is analogous to what we used when establishing (EC.13).

Now, suppose $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$ and $0 \in \mathrm{cl}(Y_p)$. We argue that

$$\nabla f(x^*)^\top (x - x^*) > 0, \text{ for } x \in F_p. \tag{EC.19}$$

First, a direct computation yields

$$\nabla g_p(0) = [B^\top\, 0]\, \nabla f_p(B\, y + x_S^*; 0)\big|_{y=0} = [B^\top\, 0]\, \nabla f_p(x_S^*; 0) = [B^\top\, 0]\, \nabla f_p(x^*) = B^\top [\nabla f_p(x^*)]_S. \tag{EC.20}$$

Since $[\nabla f_p(x^*)]_S \notin \mathcal{L}(S)$, we have $\nabla g_p(0) \neq 0$. Otherwise, due to (EC.20), $[\nabla f_p(x_S^*)]_S$ belongs to the null space of $B^\top$, which is exactly $\mathcal{L}(S)$, giving rise to a contradiction. For any $x \in F_p$, $A_S\,(x_S - x_S^*) = A_S\, x_S - A_S\, x_S^* = 0$, thus $x_S - x_S^* \in \mathrm{Null}(A_S)$. Moreover, recall that the columns of B span the whole $\mathrm{Null}(A_S)$; then there exists a $\theta \neq 0$ such that $x_S - x_S^* = B\,\theta$. Now, let $\hat{\alpha}_p = (B(B^\top B)^{-1} B^\top \nabla f_p(x_S^*); \gamma_p)$ with any $\gamma_p \in \mathbb{R}^{n-|S|}$. According to (EC.13), we have

$$\begin{aligned}
\nabla f(x^*)^\top (x - x^*) &= (\nabla f(x^*) - \hat{\alpha}_p)^\top (x - x^*) + \hat{\alpha}_p^\top (x - x^*) \\
&= \left([\nabla f(x^*)]_S - B(B^\top B)^{-1} B^\top\, [\nabla f_p(x^*)]_S\right)^\top (x_S - x_S^*) + \hat{\alpha}_p^\top (x - x^*) \\
&= [\nabla f_p(x^*)]_S^\top \left(I - B(B^\top B)^{-1} B^\top\right) B\,\theta + \hat{\alpha}_p^\top (x - x^*) \\
&= \hat{\alpha}_p^\top (x - x^*) > 0.
\end{aligned}$$

Thus, (EC.19) holds, completing the proof of Claim EC.1. $\square$

EC.8. Proof of Theorem 4

In (4), the 0-th moment constraint that $\sum_{j=1}^n x_j = 1$, together with constraint (4d), guarantees the feasible region without constraint (4c) is compact. Further, constraint (4c) can be rewritten as

$$\frac{x_j}{\sum_{k=j} x_k} - \frac{x_{j+1}}{\sum_{k=j+1} x_k} \leq 0 \text{ for } j \in [1, n-1],$$

where the LHS is continuous w.r.t. x. This gives us that the set where (4c) holds is closed, thus the feasible region is compact. By Theorem 1, problem (4) has an extreme point solution x^*. This extreme point solution corresponds to an extreme point solution y^* to problem (5) by equivalence.

Note that by constraint (4c), x^* must have consecutive support. Otherwise, assume there exist $1 \le j_1 < j_2 < j_3 \le n$ such that $x^*_{j_1} > 0, x^*_{j_3} > 0, x^*_{j_2} = 0$; however,

$$\frac{x^*_{j_1}}{\sum_{k=j_1}^n x^*_k} > \frac{x^*_{j_2}}{\sum_{k=j_2}^n x^*_k} = 0,$$

indicating x^* is not feasible. Without loss of generality, we can assume the support of x^* is $[1,n]$; that is, constraints (4d) are strict.

Let $\underline{x} := \min\{x^*_j : j \in [1,n]\}$ and define the following problem:

$$\max_{y \in \mathbb{R}^n} \sum_{j=1}^{n-1} f_j(y_j - y_{j+1}) + f_n y_n \tag{EC.21a}$$

$$\text{s.t.} \sum_{j=1}^{n} (w^i_j - w^i_{j-1}) y_j = q_i \quad \text{for } i \in [0,m] \tag{EC.21b}$$

$$y_{j-1} y_{j+1} \le y_j^2 \quad \text{for } j \in (1,n) \tag{EC.21c}$$

$$y_j - y_{j+1} \ge \underline{x}/2 \quad \text{for } j \in [1,n] \tag{EC.21d}$$

$$y_j \ge 0 \text{ for } j \in [1,n]. \tag{EC.21e}$$

The problem above is a restriction of (5) with given support, and plus (redundant) non-negativity constraint (EC.21e) so that it fits our reverse convex programming frameworks. Since y^* is an extreme point solution to (5), it is also an extreme point solution to (EC.21). Thus, we only need to verify the conditions of Theorem 2.

Again, the 0-th order moment constraint guarantees the feasible region is compact. As argued in Section 2.2, the constraint (EC.21c) is reverse convex relative to the nonnegative orthant, and so is the linear constraint (EC.21d), which implies all of the conditions in Theorem 2 are satisfied when applied to (EC.21).

At the extreme point solution y^*, the constraints (EC.21d) and (EC.21e) cannot be tight since $x^*_j = y^*_j - y^*_{j+1} > \underline{x}/2$ and y^*_j are all positive since the x^*_j are positive. Applying Theorem 2, we have that at least $n-m-1$ of the (EC.21c) constraints are tight at y^*, or, equivalently, there are at most $m-1$ of the (EC.21c) constraints that are not tight at y^*. These non-tight indexes can divide the interval $[1,n]$ into at most m pieces, and within each piece, we have

$$y_{j-1} y_{j+1} = y_j^2, \text{ for } j \in (u_i, v_i),$$

where u_i and v_i are the left and right endpoints of piece i of the domain. Thus, letting $\alpha_i = y_{u_i}$ and $r_i = y_{u_i+1}/y_{u_i}$ yields the solution (13). $\square$

EC.9. Proof of Theorem 5

In (6), the zeroth moment constraint that $\sum_{j=1}^n x_j = 1$ together with constraint (6d) guarantees the feasible region without constraint (6c) is compact. Further, constraint (6c) can be rewritten as

$$j\frac{x_j}{\sum_{k=j} x_k} - (j+1)\frac{x_{j+1}}{\sum_{k=j+1} x_k} \leq 0 \text{ for } j \in [1, n-1],$$

where the LHS is continuous w.r.t. x. This gives us that the set where (6c) holds is closed, thus the feasible region is compact. By Theorem 1, problem (6) has an extreme point solution x^*. This extreme point solution corresponds to an extreme point solution y^* to problem (7) by equivalence.

Note that by constraint (6c), x^* must have consecutive support. Otherwise assume there exist $1 \leq j_1 < j_2 < j_3 \leq n$ such that $x^*_{j_1} > 0, x^*_{j_3} > 0, x^*_{j_2} = 0$, however

$$j_1\frac{x^*_{j_1}}{\sum_{k=j_1}^n x^*_k} > j_2\frac{x^*_{j_2}}{\sum_{k=j_2}^n x^*_k} = 0,$$

indicating x^* is not feasible. Without loss of generality, we can assume the support of x^* is $[1, n]$, i.e. constraints (6d) are strict (otherwise we can redefine on the support $[k, \ell]$ and use analogous arguments).

Let $\underline{x} := \min\{x^*_j : j \in [1, n]\}$ and define the following problem:

$$\max_{y\in\mathbb{R}^n} \sum_{j=1}^{n-1} f_j(y_j - y_{j+1}) + f_n y_n \tag{EC.22a}$$

$$\text{s.t.} \sum_{j=1}^{n}(w^i_j - w^i_{j-1})y_j = q_i \quad \text{for } i \in [0, m] \tag{EC.22b}$$

$$\left\|\begin{pmatrix} j\sqrt{j-1}y_{j+1} \\ \sqrt{j-1}y_j + \frac{y_{j-1}}{2\sqrt{j-1}}\end{pmatrix}\right\| \geq j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}} \quad \text{for } j \in (1, n) \tag{EC.22c}$$

$$y_j - y_{j+1} \geq \underline{x}/2 \quad \text{for } j \in [1, n] \tag{EC.22d}$$

$$y_j \geq 0 \text{ for } j \in [1, n]. \tag{EC.22e}$$

Note that the problem above is a restriction of (7) with given support, and plus (redundant) negativity constraint (EC.22e) so that it fits our reverse convex programming frameworks.

Since y^* is an extreme point solution to (7), it is also an extreme point solution to (EC.22). Thus, we only need to verify the conditions of Theorem 2.

Again, the zeroth order moment constraint guarantees the feasible region is compact. As argued in Section 2.3, the constraint (EC.22c) is reverse convex relative to the nonnegative orthant, and so is the linear constraint (EC.22d). This implies that all of the conditions in Theorem 2 are satisfied when applied to (EC.22).

At the extreme point solution y^*, the constraints (EC.22d) and (EC.22e) cannot be tight since $x_j^* = y_j^* - y_{j+1}^* > \underline{x}/2$ and y_j^* are all positive since the x_j^* are positive. Applying Theorem 2, we have that at least $n-m-1$ of the (EC.22c) constraints are tight at y^*, or equivalently there are at most $m-1$ of the (EC.22c) constraints that are not tight at y^*. These non-tight indexes can divide the interval $[1,n]$ into at most m pieces, and within each piece, we have

$$\left\| \begin{pmatrix} j\sqrt{j-1}y_{j+1} \\ \sqrt{j-1}y_j + \frac{y_{j-1}}{2\sqrt{j-1}} \end{pmatrix} \right\| = j\sqrt{j-1}y_{j+1} + \frac{y_{j-1}}{2\sqrt{j-1}}, \text{ for } j \in (u_i, v_i)$$

where u_i and v_i are the left and right endpoints of piece i of the domain. Note the equalities above are equivalent to

$$(j-1)\frac{y_{j-1} - y_j}{y_{j-1}} = j\frac{y_j - y_{j+1}}{y_j}, \text{ for } j \in (u_i, v_i).$$

Letting

$$r_i := j\frac{y_j - y_{j+1}}{y_j}, \text{ for } j \in (u_i, v_i),$$

we have

$$y_j = y_{j-1}(1 - \frac{r_i}{j-1}), \text{ for } j \in (u_i, v_i].$$

using the fact that $j\frac{y_j - y_{j+1}}{y_j} = r$ is equivalent condition $y_{j+1} = y_j(1 - \frac{r}{j})$. Thus, letting $\alpha_i = y_{u_i}$ yields the solution (14). $\square$

EC.10. Proof of Lemma 6

If the a_j are all nonnegative or nonpositive, then $\phi(z)$ is monotone and has at most one root. Otherwise, there is an m such that $a_j \leq 0$ when $j \leq m$ and $a_j \geq 0$ when $j > m$. Denote $\phi_1(z) := -\sum_{j=1}^{m} a_j z^{i_j}$ and $\phi_2(z) = \sum_{j=m+1}^{M} a_j z^{i_j}$. Obviously, $\phi(z) = \phi_2(z) - \phi_1(z)$. Suppose $z_0 > 0$ is a root of $\phi(z)$, from non-negativity of ϕ_1 and ϕ_2 we have $\phi_2(z_0) = \phi_1(z_0) \neq 0$. Given any $z_1 > z_0$, due to (23) we have that

$$\phi_2(z_1) = \sum_{j=m+1}^{M} a_j \left(\frac{z_1}{z_0}\right)^{i_j} (z_0)^{i_j} \geq \left(\frac{z_1}{z_0}\right)^{i_{m+1}} \phi_2(z_0)$$
$$\phi_1(z_1) = \sum_{j=1}^{m} -a_j \left(\frac{z_1}{z_0}\right)^{i_j} (z_0)^{i_j} \leq \left(\frac{z_1}{z_0}\right)^{i_m} \phi_1(z_0).$$

Combining these two inequalities yields

$$\phi(z_1) \geq \left(\frac{z_1}{z_0}\right)^{i_{m+1}} \phi_2(z_0) - \left(\frac{z_1}{z_0}\right)^{i_m} \phi_1(z_0) > \left(\frac{z_1}{z_0}\right)^{i_m} (\phi_2(z_0) - \phi_1(z_0)) = 0. \tag{EC.23}$$

Similarly, for any $z_2 < z_0$, it holds that $\phi(z_2) < \left(\frac{z_2}{z_0}\right)^{i_m} (\phi_2(z_0) - \phi_1(z_0)) = 0$.

Consequently, z_0 is the only root. Moreover, when z_0 is not a root and satisfies $\phi(z_0) \geq 0$, then according to (EC.23) $\phi(z_1) > \left(\frac{z_1}{z_0}\right)^{i_m} (\phi_2(z_0) - \phi_1(z_0)) > \phi_2(z_0) - \phi_1(z_0) = \phi(z_0)$ implying that $\phi(z)$ is monotonically increasing on $\{z \mid \phi(z) \geq 0\}$. $\square$

EC.11. Proof of Lemma 7

It suffices to provide a uniform bound on $|\frac{\partial g_2(\alpha,\beta)}{\partial \beta}|/|\frac{\partial g_1(\alpha,\beta)}{\partial \beta}|$. When $a \leq 0$, since the distribution has positive mass on both sides of a, it has positive mass at $X < a$. This combined with (24) implies that $\mathbb{E}[X \cdot (a - X)\mathbb{1}_{X<a}] \leq \mathbb{E}[-(X-a)^2 \mathbb{1}_{X<a}] < 0$. By further invoking (26) and (22), we have that

$$\begin{aligned}
\frac{|\frac{\partial g_2(\alpha,\beta)}{\partial \beta}|}{|\frac{\partial g_1(\alpha,\beta)}{\partial \beta}|} = \frac{|\mathbb{E}[(1-X^2)(a-X)\mathbb{1}_{X<a}]|}{|\mathbb{E}[X \cdot (a-X)\mathbb{1}_{X<a}]|} &\leq \frac{|\mathbb{E}[(1-X^2)(a-X)\mathbb{1}_{X<a}]|}{\mathbb{E}[(X-a)^2 \mathbb{1}_{X<a}]} \\
&\leq \max_{x: x=a-j\epsilon, j=1,2,\cdots,\tilde{k}} \frac{|(1-x^2)(a-x)|}{(x-a)^2} \\
&\leq \frac{\max_{x: x=a-j\epsilon, j=1,2,\cdots,\tilde{k}} |(1-x^2)|}{\min_{x: x=a-j\epsilon, j=1,2,\cdots,\tilde{k}} (a-x)} \\
&\leq \frac{1+D^2}{\epsilon},
\end{aligned}$$

where the second inequality is due to the fractional linear function $\frac{y}{z}$ is quasi-convex and thus $\frac{\sum_i \alpha_i y_i}{\sum_i \alpha_i z_i} \leq \max_i\{\frac{y_i}{z_i}\}$ for $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$.
Similarly, when $a > 0$ and $\mathbb{E}X = 0$ implies that $\mathbb{E}[X \cdot \mathbb{1}_{X<a}] + \mathbb{E}[X \cdot \mathbb{1}_{X\geq a}] = 0$. Therefore,

$$a\,\mathbb{E}[X \cdot \mathbb{1}_{X<a}] = -a\mathbb{E}[X \cdot \mathbb{1}_{X\geq a}] \leq -a^2 P(X \geq a),$$

and due to (24)

$$\mathbb{E}[X \cdot (a-X)\mathbb{1}_{X<a}] \leq \mathbb{E}\left[-X^2\,\mathbb{1}_{X<a}\right] - a^2 P(X \geq a) \leq \mathbb{E}\left[\min(X,a)^2\right] < 0,$$

where the last inequality follows from the fact that X has positive mass at both a and $X < a$. Since the domain of X is bounded with diameter D, we also have that for any $b > 0$

$$\begin{aligned} 1 = \mathbb{E}X^2 = \mathbb{E}[X^2\mathbb{1}_{|X|<b}] + \mathbb{E}[X^2\mathbb{1}_{|X|\geq b}] &\leq b^2\mathbb{E}[\mathbb{1}_{|X|<b}] + D^2\mathbb{E}[\mathbb{1}_{|X|\geq b}] \\ &= b^2(1 - P(|X| \geq b)) + D^2 P(|X| \geq b). \end{aligned}$$

That is $P(|X| \geq b) \geq \frac{1-b^2}{D^2-b^2}$. Furthermore, note that for any $|X| \geq b$ we have $|\min(X,a)| \geq \min(b,a)$. Then,

$$\begin{aligned} \mathbb{E}\left[\min(X,a)^2\right] &= \mathbb{E}\left[\min(X,a)^2 \cdot \mathbb{1}_{|X|\geq b}\right] + \mathbb{E}\left[\min(X,a)^2 \cdot \mathbb{1}_{|X|<b}\right] \\ &\geq \mathbb{E}\left[\min(X,a)^2 \cdot \mathbb{1}_{|X|\geq b}\right] \\ &\geq \min(b,a)^2 P(|X| \geq b) \\ &\geq \min(b,a)^2 \frac{1-b^2}{D^2-b^2}. \end{aligned}$$

Since $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$, the upper bound D of $|X|$ is greater than 1. Therefore, when $a > 0$ by taking $b = 0.5$, we have $D^2 - b^2 > 0$ and when $a > 0$ it holds that

$$\frac{|\frac{\partial g_2(\alpha,\beta)}{\partial \beta}|}{|\frac{\partial g_1(\alpha,\beta)}{\partial \beta}|} = \frac{|\mathbb{E}[(1-X^2)(a-X)\mathbb{1}_{X<a}]|}{|\mathbb{E}[X \cdot (a-X)\mathbb{1}_{X<a}]|} \leq \frac{|\mathbb{E}[(1-X^2)(a-X)\mathbb{1}_{X<a}]|}{\mathbb{E}[\min(X,a)^2]} \leq \frac{(1+D^2)(a+D)}{\min(0.5,a)^2\frac{0.75}{D^2-0.25}}.$$

Combining the bound on $a \leq 0$, we conclude that

$$\frac{|\frac{\partial g_2(\alpha,\beta)}{\partial \beta}|}{|\frac{\partial g_1(\alpha,\beta)}{\partial \beta}|} \leq \begin{cases} \frac{(1+D^2)(a+D)(D^2-0.25)}{(0.75)\min(0.5,a)^2}, & a > 0 \\ \frac{1+D^2}{\epsilon}, & a \leq 0 \end{cases}.$$

That is, the absolute value of $|\frac{\delta}{\delta\beta}g_1(\alpha,\beta)|$ is uniformly bounded below. $\square$

EC.12. Proof of Theorem 7

Since ϕ' is a continuous function, there exists $\delta > 0$ such that $|\phi'(z) - \phi'(\gamma)| \le \frac{1}{4}|\phi'(\gamma)|$ whenever $|z-\gamma| \le \delta$. Consider a Newton step with $|z_k - \gamma| \le \delta$ and

$$z_{k+1} - \gamma = z_k - \gamma - \frac{\phi(z_k) - \phi(\gamma)}{\phi'(z_k)} = \frac{\phi(\gamma) - \phi(z_k) - \phi'(z_k)(\gamma - z_k)}{\phi'(z_k)} = \frac{(\phi'(\gamma_k) - \phi'(z_k))(\gamma - z_k)}{\phi'(z_k)},$$

where γ_k is a point between γ and z_k. Note that $|\phi'(\gamma_k) - \phi'(z_k)| = |\phi'(\gamma_k) - \phi'(\gamma) + \phi'(\gamma) - \phi'(z_k)| \le \frac{1}{2}|\phi'(\gamma)|$ and $|\phi'(z_k)| \ge \frac{3}{4}|\phi'(\gamma)|$. Therefore,

$$|z_{k+1} - \gamma| \le \frac{|(\phi'(\gamma_k) - \phi'(z_k)| \cdot |\gamma - z_k|}{|\phi'(z_k)|} \le \frac{2}{3}|z_k - \gamma| \le \delta.$$

Moreover, we have

$$\lim_{k\to\infty} \frac{|z_{k+1} - \gamma|}{|z_k - \gamma|} = \lim_{k\to\infty} \frac{|\phi'(\gamma_k) - \phi'(z_k)|}{|\phi'(z_k)|} = 0,$$

and we conclude that $\{z_k\}$ is superlinear convergent. $\square$

EC.13. Proof of Theorem 8

According to (25) and (21), we have that $|\frac{\partial g_2(\alpha,\beta)}{\partial\alpha}| \le \frac{(1+D^2)(a+D)}{\rho\cdot\epsilon}$ and $|\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}| \le \frac{(a+D)D}{\rho\cdot\epsilon}$ respectively. Finally, by invoking (27) we obtain an upper bound of gradient of $g_2(\cdot, h(\cdot))$:

$$\begin{aligned}
\left|\frac{\mathrm{d}g_2(\alpha, h(\alpha))}{\mathrm{d}\alpha}\right| &= \left|\frac{\partial g_2(\alpha,\beta)}{\partial\alpha} + \frac{\partial g_2(\alpha,\beta)}{\partial\beta} \cdot \frac{-\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}}{\frac{\partial g_1(\alpha,\beta)}{\partial\beta}}\right| \\
&\le \left|\frac{\partial g_2(\alpha,\beta)}{\partial\alpha}\right| + \left|\frac{\partial g_2(\alpha,\beta)}{\partial\beta}\right| \cdot \left|\frac{-\frac{\partial g_1(\alpha,\beta)}{\partial\alpha}}{\frac{\partial g_1(\alpha,\beta)}{\partial\beta}}\right| \\
&\le \begin{cases} \frac{(1+D^2)(a+D)}{\rho\epsilon}\left(1 + \frac{D(a+D)(D^2-0.25)}{(0.75)\min(0.5,a)^2}\right), & a > 0 \\ \frac{(1+D^2)(a+D)}{\rho\epsilon}\left(1 + \frac{D}{\epsilon}\right), & a \le 0 \end{cases}.
\end{aligned}$$

$\square$